第一章　概论

当今世界，科学技术的发展日新月异。其中，生命科学的进展尤为引人注目。进入分子水平以来，人们发现在生物化学、分子生物学、免疫学以及遗传学领域的研究中有大量的数据资料需要处理。于是，随着计算机技术、网络通讯的飞速发展，产生了一门新兴的学科——生物信息学。它首先利用电子计算机技术，对在分子生物学等学科的研究中产生出来的大量原始数据进行收集、整理和管理；其次，对各种数据进行对比、分析、归纳并建立计算模型，以期更好地解释数据，并进行结构、功能的预测以及仿真，等等(图 1-1)。它的出现极大地推动了分子生物学的发展，在人类基因组计划的研究中发挥了重要的作用。这门学科在生物学、医学领域有着十分广泛的应用。其中的一些大型生物学数据库包含了众多的生物学信息资源，人们可以很方便地从国际互联网上寻找

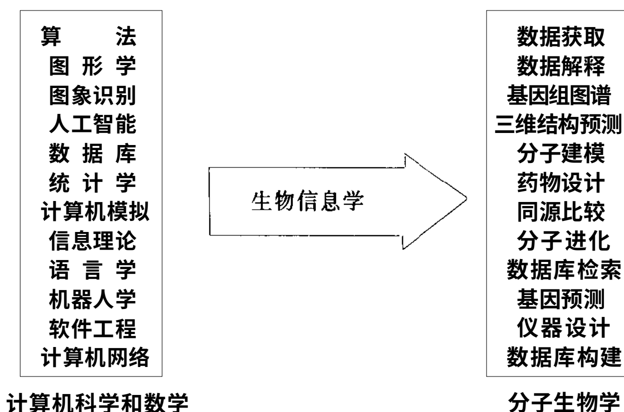


图 1-1　生物信息学是计算机科学、数学和分子生物学之间的桥梁

所需的资料和处理工具。这不仅方便了研究思想和资料的交流，减少了许多重复性的工作，而且也提供了一种崭新的思维方式和科研工作方法。近年来，互联网的高速发展为人们共享数据资源、合作研究提供了网络这一物质基础。越来越多的生物学、医学、药学工作者认识到生物信息学的重要性和实用性，其良好的发展前景业已显现。

第一节　生物信息学及其与生物学的关系

近十多年来，生命科学在分子水平上进行了广泛而深入地研究。随之而来的是大量的数据结果需要处理。特别是生物化学、分子生物学及遗传学的研究，各种各样的有关生物分子的原始实验数据,数量十分庞大。因此,利用计算机技术处理数据十分必要。另外,众多的学科,诸如结构生物学、酶学、细胞生物学、生理学、病理学、神经生物学等等，从不同角度的研究结果，可经过计算机的分类、组织和构建，形成具有生物学意义的新的研究结果。这些新的结果是对生命体的细胞结构和功能更为本质的反映。在这样的情形下，生物信息学应运而生。

生物信息学(Bioinformatics)的萌生可以追溯到 1956 年,那时还是计算机的初创期。在美国田纳西州的 Gatlinburg,曾召开过首次“生物学中的信息理论讨论会”，这拉开了生物信息学的序幕。随着二十世纪八、九十年代计算机技术的迅猛发展，它才同时获得自身的快速成长。无论从理论上来讲，还是从现实情况来看，生物信息学都还是一门相当年轻的学科，它的实质就是利用计算机科学和网络技术来解决生物学问题。它的诞生和发展是应时所需，是历史的必然，并且已经悄然渗透到生命科学的每一个角落。以至于在整个科学界意识到它的存在之前，相关学科的研究者就已经离不开它了。

二十世纪末期，生命科学技术的迅猛发展，无论从数量上还是在质量上，都极大地丰富了生命科学的数据资源。数据资源的急剧膨胀首先迫使人们不得不考虑寻求一种强有力的工具，在有效地组织数据的同时，有利于对已知生物学知识的储存和进一步地加工利用。在大量多样化的生物学数据资源中，必然蕴含着许多重要的生物学规律。这些规律是我们解决许多生命之谜的关键所在。然而，继续沿用传统手段以人脑来分析如此庞杂的数据是不可能的。人们同样需要寻求一种强有力的工具去协助人脑完成这些分析工作。可以说，伴随着二十一世纪的到来，生命科学的重点和潜在的突破点已经由上个世纪的试验分析和数据积累，转移到数据分析及其指导下的实验验证上来。生命科学也正在经历着一个从分析还原思维到系统整合思维的转变。

那么，我们所寻求的那种强有力的数据处理分析工具，就成为未来生命科学的关键所在；伴随着生命科学这一需求的加剧，以数据处理分析为本质的计算机科学技术和网络技术获得了突飞猛进的发展，而自然地成为生命科学家的必然选择。计算机科学技术和网络技术正日益渗透到生命科学的方方面面，一门崭新的、拥有巨大发展潜力的生物信息学也就悄然而坚定地发展起来了。可以说，历史必然性地选择了生物信息学——生命科学与计算科学的融合体——作为新一代生物科学研究的重要工具。

生物信息学(Bioinformatics)这一名词的由来，还要从八十年代末期说起。美国佛罗里达州立大学超级计算机计算研究所的林华安博士认识到将计算机科学与生物学结合起来的重要意义，遂开始留意为这一新的领域构思一个合适的名称。考虑到与佛罗里达州立大学大型计算机计算研究所的关系，起初，他使用的是“CompBio”。当时,这一机构支持由他主办的一系列“生物信息学”的会议；之后，他又将其改为兼具法国风情的“bioinformatique”。因其拼写看起来似乎有些古怪，不久，他便进一步把它更改为“bio-informatics(或 bio/informatics)”但由于当时的电子邮件系

统与今日不同,该名称中的‘一’或‘／’符号经常会引起许多系统问题。于是，林博士又将其去除。今天，我们所看到的“bioinformatics”就这样正式诞生了。林华安博士也因此赢得了“生物信息学之父”的美誉。

一、生物信息学的定义

生物信息学主要是由分子生物学与信息学、计算机技术、数学、物理学等学科交叉结合的产物。对于这样一门年轻的边缘科学,不同的学者对它的定义不尽相同(见附录二)。有不严格的定义称之为：分子生物学与计算生物学的交叉学科。国外学者一般认为，它是对现代分子生物学和生物化学技术带来的不断增加的复杂的资料进行分析、组织并使之系统化的一门科学。也有人认为，生物信息学应含有生物系统内信息链的内容，它主要指的是贮存于 DNA 或 RNA 中的信息，表现为核苷酸的序列并能通过翻译表达出重要的生命大分子——蛋白质。对这部分内容的研究,无疑是生物信息学在应用上的一个很重要的方面。我们认为生物信息学的含义是基于计算机和互联网的应用和信息科学的知识方法对生物信息进行收集、整理、分析研究、处理和应用的一门交叉学科。今天已经认识到：一项研究欲更加深刻地反映生物的本质规律，需要用到这门新兴的学科。例如，基因密码的含义与相对应的生物机体生理特点之间的关系、人脑的研究、基因与意识及心理行为的关系、系统遗传学家对各物种之间内在关系的研究等等，这类研究均需在计算机软件技术、各种不同类型的生物学数据库的辅助下完成。又如，结构生物信息学对靶蛋白质活性位点精细结构的描述可为新药的模拟设计提供良好的基础。总之，生物系统的复杂性需要生物学方法与计算技术的结合。所以，生物信息学是一门建立、管理并运用生物信息数据库研究生命现象，并最终模拟出生命有机体复杂性的科学。

一门学科的建立除了有应用上的需求外，还应当有相应的理论支持。信息学理论的发展是其重要的支柱之一。此外，计算机凭

借其强大的运算分析功能介入到生物学的研究中，使研究手段、工具方法迈上了新台阶。美国学者 H. Rashidi 和 L.K. Buehler 就认为生物信息学是建立在这样一个假设的基础上的：即基因结构、基因在基因组中的排列位置、蛋白质的功能以及在机体中引起能量代谢、繁殖和构成诸如身材、体型等蛋白质的相互作用之间存在着一个分等级的关系。而对其相互关联的研究，使人们意识到计算方法的介入为此提供了一个良好的平台。

二、生物学的发展与生物信息学

二十世纪初，人们用有机化学的方法研究三大物质的代谢途径，研究酶的组成及生理作用，等等。那时候，生物化学家没有分子生物学、基因的知识，并不知道核酸是生命的遗传单位。他们的研究是对各种实验现象的观测和记录。而时至今日，人们已经可以在电脑前完成基因测序、基因筛选、计算机识别蛋白质功能、计算机模拟蛋白质三维结构以及新药设计等工作，发展出计算和实验方法相结合的新的生物学研究模式。下面试举一例，来说明生物信息学在这一新模式中的用途。

基因是生命的遗传单位。在复制时，保持基因中分子信息的严密性和准确性是十分重要的。我们在研究某个基因突变与肿瘤发生的关系时，该基因的克隆是首先应完成的，因为这是获得核酸序列及寻找调节因子的第一步。首先，将我们需要的 DNA 片段从有关的基因组中分离出来，然后将这段基因插入到一个载体 DNA 中,从而制成重组 DNA。按生物进化的观点,所有生命体在遗传上是有密切的相关性的，所以人类基因在其他动物体或微生物体内操纵复制是完全有可能的。由此，人们将上述重组 DNA 置入细菌体内繁殖，从而达到基因克隆的目的并可复制出大量的基因拷贝。应用这种方法复制基因、扩增 DNA 简单而有效,而且避免了为纯化 DNA 或蛋白质而需获取大量的人体组织标本的过程。

基因克隆完成后，即可对基因测序。通过基因序列可预测其

相应蛋白质的结构和功能。这些工作如今已可在计算机辅助下完成。而重组 DNA 又可用来合成相应的蛋白质，对后者进行生物化学的检测分析，以进一步明确其结构、功能及在致病过程中的作用。对肿瘤基因及遗传性疾病的研究中，为了明确该致病基因在基因组中的定位，常常需要获得携带有突变基因的个体样本及正常人的样本。在实践中，对一定数量的两种样本的对比分析可运用相应的计算工具。这种工具是按医学研究的目的而建立起来的生物学数据库，并经过不断地调整编辑而成。毫无疑问，这一编辑过程也促进了人们对疾病本质及其遗传本质的理解。

二十世纪八十年代后期，计算机技术进入快速发展时期，此后的互联网以更高的速度在全球铺展开来。与此同时，一项庞大的人类基因组计划及其他的生命体基因组研究业已全面展开。这些均是在生物信息学形成和发展中具有决定性的事件。在基因组计划中，人们更关注基因的核酸序列。在获取基因序列并揭示其中的生物信息的过程中，生物信息学是重要的分析工具。

当前，分子生物学与生物信息学的结合愈发紧密。后者为前者提供了新的研究手段和方法，如今已在基因克隆、核酸测序、基因定位等方面有着广泛地应用。在人类基因组研究及后基因组的研究工作中，效率是经常会被提及的因素之一。在 DNA 的序列研究中，任何一种计算方法都比实验分析要迅速和廉价，且计算分析为实验分析提供了互补的预测性信息。现有的算法在利用已知的生物学知识的基础上，已经完成了不少工作。可以预见：在未来，对生物学有了更深入的研究之后，计算分析学家和实验生物学家会有更频繁更深入的合作，这一领域会有更为显著的进步。以计算和实验相结合的新生物学，已快步向我们走来。

三、基因组学、蛋白质组学与生物信息学

如今，基因组学、蛋白质组学已是生命科学研究中最重要的内容之一。传统上，要获得一个基因序列，需要 mRNA 的分离或检测蛋白质的氨基酸序列。其后，通过诸如蛋白电泳等检测手段，

探寻该基因及其相关蛋白质与生物体的发生、发育、老化及疾病发生之间的联系。基因的组成、表达等与生物体的生理功能相关的信息，在阅读理解基因图谱时，有重要的意义。现代基因组计划有两方面的工作：其一，基因组结构是与一定的生理功能相关的。所以，人们希望通过研究一个生命体的全部基因组序列，以帮助了解其生物学特点；其二，高等生物含有大量的非编码 DNA，直到最近人们才对其功能及存在的意义有了初步的了解。在过去，人们采用功能性的检测方法不能获得非编码 DNA 的序列信息，而未来的研究会进一步加深对它的认识。

科学家的研究是从 DNA 开始的，但在生物体内它只是遗传信息的载体。因为 DNA 在体外是不能自我复制的，所以核酸并不是单独完成遗传使命的，DNA 上所携带的遗传信息必须首先被解读。在细胞中，这一工作是由一些蛋白质承担的，诸如存在于胞核或胞浆中的蛋白质及酶等等，都是解读遗传信息的工具，而且在胚胎发生的早期就起着重要的作用。尽管基因成对出现在染色体上，但表达的只有一条，它来自父方或源于母方。所以，并不仅仅只有 DNA 序列是遗传信息。染色体的结构， DNA 与其表达的蛋白质间的相互关系以及其构型组合也是信息的一部分。一种影响或决定子代中母系抑或是父系的基因激活的表达机制叫遗传印迹（Genetic Imprinting）。要搞清楚这一现象及其基因剂量效应（一个基因能表达多少蛋白质是与其相关基因有联系的），就必须研究发生在细胞内的整个遗传过程的时空调控。因此，有人认为人类基因组计划一旦完成，接下来的工作将由分子生物学实验室转入由电子仪器组成的实验室。在那里，基因的各种信息会轻松地获得，这有利于确定基因在发育、衰老及疾病发生等过程中的作用。

蛋白质组学是基因组计划完成后或同时开展的重要研究领域之一。要明确各种器官、组织、细胞以及正常和疾病组织中多种蛋白质表达谱的变化，即蛋白质的大规模识别和定性，需要有强

有力的分析工具。2D凝胶电泳是十分有用的方法，在2D凝胶电泳的结果分析过程中离不开生物信息学，后者使这一分析过程自动化。本书有专门章节介绍生物信息学在蛋白组学中的应用。

当前，数据库内容的增加及变化都十分迅速和频繁。因为每天都可能有新提交的数据加入其中，所以数据库的目录可能每天都在更新。这更有利于研究机构的获取和利用。有资料显示：在1998年4月包括83个物种的基因组计划已完成了21个物种的测序，其中大部分为微生物。这一工作采用了自动克隆和聚合酶链式反应（PCR）完成DNA的扩增及测序。这些方法可以把由盲法产生的随机DNA片段重建为无间沟（gap）的连续序列，并最终得到全部基因组的所有碱基序列。在进行基因组计划的过程中，每天都有大量的信息出现，并进入到数据库中。

基因组研究院(The Institute for Genomic Research，TIGR)创建于1992年，是一个非赢利性的研究院。它位于美国马里兰州的Rockville，与美国国立卫生院(NIH)、约翰一霍普金斯大学、马里兰大学及其它研究所、生物技术公司毗邻。占地12公顷，有50000平方英尺的实验室及办公区域。该研究院有大型的DNA测序实验室和与生物信息学、生物化学和分子生物学相关的现代化设备。它从一开始就利用网络成长起来。对于许多科学家来说，TIGR使他们开始真正认识了基因组计划，并获得了一种新的大批量测序的方法。TIGR的工作促进了测序程序及数据分析的发展。类似TIGR这样的研究院和组织还有一些，他们在网络上提供的信息数量之大令人惊讶。TIGR的研究对象是病毒、真菌、致病菌、原生质、真核生物以及人类的基因组，并对基因产物的功能、结构进行比较分析（摘自 http://www.tigr.org/about/）。

TIGR率先发展了大量复制DNA的必需技术，以及EST(Expressed Sequence Tag，表达序列标签）测序计划。EST文件提交格式简单、易于快速处理，因此每天提交进入数据库的数量级可以达到数千个，峰值期可达每周100，000个提交量。这种测

序形式已在一定程度上影响了生命科学界的工作方式。因为现今大多数的期刊已不再刊登完整的序列数据，而只标明其序列在数据库中的序号。研究者在公开发表文章时要向公共数据库提交其研究结果，这已成为一条准则。而且，有些大型基因研究中心规定新发现序列的公开应先于论文的发表。这些情况都使得相关数据库的内容呈指数级上升，同时也使数据信息的整理、分析、利用显得十分重要。

最初，生物信息学就是为来自不同国家、不同研究组织之间的信息交流服务的，是他们相互合作、信息共享的方式之一。随着数据库的集中合并及网络交流的迅猛发展，它不仅成为了业内人士主要的交流方式，而且很快转变为一门独立的学科。特别是人类基因组计划的实施，更多的愈来愈强大的克隆和测序技术不断出现，直接促进了生物信息学的发展。在这项庞大的计划中，国际间的合作成为了必然。由此而来，出现了拥有各种数据库的公立或私立的组织，他们的数据库为整个基因组测序、基因定位以及在细胞或分子水平上寻找 DNA 序列信息与结构功能的关系提供了实用的工具和便捷的服务。

其后，工商企业界及金融投资者逐渐认识到生物信息的处理和出售极具潜在的利润价值。他们的介入使得数据库的建立、完善有了充足的资金来源，并且在其未来发展中扮演着重要的角色。潜在的利润和商机，又促进了各种基因研究工作的深入并使其竞争日趋激烈。

四、国内生物信息学现状及展望

国际上，欧美等国家在生物信息学的研究和应用方面已经有了较长时间的积累。国内对生物信息学领域也越来越重视，在一些著名院士和教授的带领下，在各自领域取得了一定的成绩。2001 年 4 月在北京召开了中国首届生物信息学大会，参会人员遍及全国 10 多个省市，共 600 余人。此次会议较为全面地回顾了我国生物信息学研究的现状。 2001 年 10 月，国家发展计划委员会宣布，

我国将在中国科学院建设"生物信息系统国家研究中心"，形成有国际竞争能力的基因组学、蛋白质组学和生物信息学的整体技术平台。这将会推进我国生物信息技术的发展。下面将国内部分单位的生物信息学发展状况做一简单的介绍。

1. 中国科学院基因组信息学中心生物信息学平台

中国科学院基因组信息学中心设有专门的生物信息室，配备有由国家智能计算机研究开发中心研制的曙光 3000 型大型计算机。这是目前国内性能最高、运算速度最快的超级服务器。该系统峰值浮点运算速度为每秒 4032 亿次，内存总量为 168GB，磁盘总容量为 3.63TB。它具有先进的体系结构，丰富而完善的软件系统和一大批行业应用软件。该生物信息学平台负责的项目包括：人类基因组计划中国部分完成图、嗜热菌基因组、螺旋藻基因组、超级杂交水稻基因组工作框架图和中华民族基因组及疾病相关基因的多态性研究等。

2. 北京大学生物信息学服务器

北京大学生物信息学服务器是在罗静初和顾孝诚教授领导下建立的，由北京大学附属的分子设计实验室和物理化学研究所维护。它是国内第一家生物信息学网站，设有多个国外著名分子生物学数据库的镜像站点，如：Protein Data Bank (PDB)、Structural Classification of Protein (SCOP)、Protein Information Resources (PIR)、SWISS-PROT、ENZYME、PROSITE、BLOCKS 等，在国内查询这些数据库亦非常方便快捷。他们开展的项目包括蛋白质结构预测、以结构为基础的药物设计、蛋白建模和设计等方向的研究。

3. 联众研究院生物信息分析平台

该生物信息分析平台隶属于上海复旦大学，他们建立了自已的 EST 数据库、Cluster 数据库（UniGene）和全长基因数据库，并从国外引进了 GenBank、SWISS-PROT、EMBL、OMIM、UniGene 等数据库。每天能开展对约 1500 个序列进行公开数据库

的查询、全长基因的识别、全长基因编码蛋白的结构与功能预测、部分全长基因的染色体定位等方面的工作。对外提供的生物信息学服务包括：引物设计、核酸一级结构、同源核酸序列数据库搜索分析、ORF 预测、氨基酸组成、理化特性的分析、蛋白质功能域分析、基因或蛋白家族分析、基因组 DNA 的外显子区域预测、ESTs 与基因组序列比较、蛋白质亲水性分析、蛋白质跨膜区预测、信号肽预测、序列抗原性分析、二级结构预测等。他们开发了中文环境的软件 Biolink，可以用于计算蛋白质等电点分析、蛋白质二级结构的分析预测、一条或多条序列在一个或多个核酸或蛋白序列库中进行同源搜索、DNA 限制性内切酶图谱分析、识别基因 ORF 编码区、预测蛋白质在细胞内定位、分析蛋白质的一些理化性质（如：亲／疏水性、跨膜片断等）。具有识别跨膜螺旋区、分析氨基酸的组成、PCR 引物和杂交探针设计等功能。

4. 中国人民解放军总医院神经信息中心

该中心成立于 2001 年 9 月。人类脑研究计划是继人类基因组计划之后又一国际性的重大科研项目，其核心是神经生物信息学。科学界认为该计划比基因组计划规模更大，囊括了更加广泛的内容。人类脑研究计划的目标是提供先进的信息学工具，使神经科学家和信息学家能够将脑的结构和功能研究结果联系起来，建立数据库，进行搜索、比较分析、合成和整合，绘制出脑功能、结构和神经网络图谱。中国人民解放军总医院神经信息中心的主要任务是：建立神经信息工作平台，为开展神经信息学研究提供必要的条件。其中包括：在国内 6 大城市 11 个研究单位开通神经信息电子网络，进行网上信息交流和科研协作；与国际神经信息电子网络接轨，引进和推广全球性“人类脑计划”的科研成果；开展神经信息科研工作，组织全国性脑研究计划；拟成立相关工作组，以代表中国加入全球神经信息学工作组织，参与全球人类计划的研究工作；提供神经信息服务。

虽然国内生物信息学的发展非常快，但总体来讲与国际水平

差距还比较大。一方面表现为相对于国内生物医药科学的研究与开发，对生物信息学的研究和服务的需求滞后；另一方面是，真正开展生物信息学服务的公司相对较少。仅有的几家科研机构主要开展生物信息学的理论研究，而声称提供生物信息学服务的公司所提供的服务也仅局限于简单的计算机辅助分子生物学实验设计，而且服务体系并不完善，这就与欧美发达国家有了较大的差距。

生物信息学的产业特点是投资少、见效快、效益大，适合我国的现实条件。如果从互联网上源源不断地采集数据，然后进行分析、归类与重组，发现新线索、新现象和新规律，用以指导实验工作的设计，这是一条既快又省的科研线路，可以避免不必要的重复，提高我国生命科学的研究水平。其关键在于加速培养一批在数学、物理、计算机科学和分子生物学方面均有造诣的跨学科青年人才。如能充分发挥现有人才的潜力，进一步培养大批生物信息学的专业人员，才能迎接21世纪的挑战。

第二节　计算机在生物学及医学领域的应用

一、生物学、医学与计算机

众所周知，技术的进步对科学的发展起到了重要的促进作用。在最近的二十年，这一趋势更加明显。例如，纳米，这一奇妙世界的物理尺度，也是生命分子本身各种组成部分的尺度。纳米技术是一种新近发展起来的对单个分子进行操作的技术，现已成为一个时髦的研究领域。该技术显示出：将以单分子机械装置为目标，促进医疗和电子技术的微型化。又如，材料科学把化学与生物化学有机地结合在一起。在生命科学领域，生物样品的准备过程中发展起来的荧光染色技术，引起了细胞生物学和DNA操作技术的革命（例如：Affymetrix公司的DNA芯片技术）可视化的脑检测技术又使脑科学的研究进入了一个新的层次。化学和生

物学研究方法的结合又开拓出一些新的研究领域（如：组织工程），给二者的发展注入了活力。同样的，计算机技术在生物医学领域也有很广泛的应用。

应用数学及计算机科学是现代生物学的重要研究工具。假如没有计算机对信息的存储和读取，没有数据装载和统计分析，没有计算机模拟系统，就不可能产生现代分子生物学。可见，计算机在这一领域发挥了重要的作用。从软件设计、 PC 机的应用到互联网的交流，都充分利用了计算机技术。而且，几乎在所有的科研活动中，它都发挥了日趋重要的作用。在未来，这一作用将更为明显。

在基础医学研究中，实验就是在特定的时间内通过一系列的技术手段检验一种观点或得到新发现的过程，其结果如何并非出自偶然，而是由事物的必然性决定的。现在认为，计算机工具在实验的设计、执行和分析研究过程中可以起到核心的作用。计算机的介入并没有改变科学思想本身，也没有改变围绕着科学发现与错误模型之间的假设与争论，但却改变了科学研究的核心内容——具体实验的本身。计算机可以计数培养皿中的细胞，测量显微镜下的各种组织切片中的细胞核大小，记录对慢性疼痛敏感的神经元的电活动，读取电泳凝胶上可记录于 光底片上的 序列。其他不可缺少计算机的实验室设备有：流式细胞仪、远程病理诊断系统、 芯片分析仪、DNA 凝胶成像系统等等。计算机可以帮助人们快速而精确地记录许多重要数据，明显提高了实验的精确性。

当然，实验的精确性并不完全取决于计算机，而主要是实验仪器的质量。如：电镜下精确切割冰冻细胞样品、把微玻璃管插入干细胞转移细胞核以培育转基因鼠、或者测量脑组织中单个神经元的电活动，都需要设计制造出高质量的合金。计算机在生命科学研究中的作用依然是控制、运算、数据分析及存储。计算机的数字化记录及其易于复制的特点，极大地提高了生物数据的存

储 量 。

临床医学是计算机技术的另一个受益者。两种成功的无创诊断技术——核磁共振成像和超声波检查，均是有效地利用了所有物质中特异性原子或分子的物理特性而成像的，计算机技术在其中起到了重要的作用。另外，计算机专家系统的发展和使用，将会使常规医疗变得愈来愈方便，且治疗的成功率也愈来愈高。在临床科研及临床诊治中，精密的检验和治疗仪器的使用都离不开计算机。现代医学借助了许多具有分析功能的仪器及新颖的实用医疗操作技术辅助医生诊治疾病。例如：用来监测血糖水平的生物传感器、用作微血管介入技术的导管等等，都与计算机技术相关。目前，国外医学领域的科研医疗与计算机的联系已十分密切，甚至已呈现出较强的依赖性。尤其有代表性的是“千年虫”问题，在 1999 年底给发达国家的医学界也带来了不少麻烦和干扰。

生物信息学是计算机与生物学紧密结合的产物。人类基因组计划旨在绘制一幅染色体图谱。也正是由于这一计划中产生的大量序列信息，大大地促进了生物信息学的发展。而神经生物学正在绘制由大脑解剖及其细胞组分构成的图谱。大脑是一个十分复杂的器官，对它的研究意义重大。但利用传统技术研究多年，进展有限。在神经生物学家、认知科学家及心理学家的携手合作下，一门新兴的边缘学科——神经生物信息学出现了。这一领域实际上也属于生物信息学的范畴。这是研究大脑与神经元结构和功能的新途径，是理解以网络形式存在的复杂的神经系统的新途径。

二、计算机算法

在预计计算机解决某一问题需要多少时间时，这里面其实包含了两部分内容：其一，计算机自身计算时所需要的时间；其二，分析指令的时间，这是人为干预占用的时间。所以，现在的许多自动程序软件以培养计算机的自主决定能力为目标，以避免人的操作占用了较多的时间。这样更有利于计算机在短时间内处理大量的数据。计算机是以其特殊的专有系统作为其运行的基础，这

一系统可完成一项或多项计算量很大的任务。过去，许多需要人们操作干预完成的工作，现在已经可由具有学习功能的神经网络（neural networks，NNs）作出正确的答案。神经网络尽管发展前景可观，但仍难以成功地解决如何对符号和记忆进行控制的问题，并且无法训练 NNs 产生数据之外的信息。一旦建立了一种计算方法，并且成功地运行后，计算机可不停地重复这一过程。并且，在不断地输入及输出的过程中，通过反馈及反馈环作用进行调整，以适应预先设定的要求。除了神经网络算法外，还有许多种著名的算法，如 HMM（Hidden Markov Model ）等等应用于生物信息学。在人类基因组计划的序列拼接中，生物信息学发挥了很重要的作用，目前 DNA 自动测序仪的每个反应只能测序大约 500bp。如何将这些序列片段拼接成完整的 DNA 顺序，就成为测序后的一项重要工作。传统的测序技术通常是将克隆进行亚克隆，并对亚克隆进行排序，这些工作需要大量的人力物力。现在，生物信息学提供了自动而高速地拼接序列的算法，即根据 Lander-Waterman 模型利用鸟枪法进行测序，再将大量随机测序的片段用计算机进行自动拼接。这种技术不仅避免了亚克隆排序所需的大量繁琐的工作，还使序列具有一定的冗余性以保证序列中每个碱基的准确性。

可见，计算机算法所表现出的优越性是无可置疑的，但对其不足也应有充分的了解。比如说，一个文字处理程序因可使书写变得轻而易举而大受欢迎，成为必要的文案书写工具。尽管编辑一个文本所需的时间不多，但是由于文本和图表排列很容易改变，反而使得纸张的浪费增多了。人们更愿意看到打印在纸上的文本，它似乎比荧光屏上的文件更可靠些，而且符合相当一批人的阅读习惯。但是，计算机的拼写监测器是缺乏语言分析能力的。其中的一个难题是如何校正文本的内容，如果输入的内容与原文不符但拼写无误，计算机就无法识别这种版面上的错误。而人脑之所以可识别这种区别，是因为它与计算机的工作机制不同。虽然校

对、分析数据以及随后的内容解释仍需依靠计算机的辅助，但这是在人脑的严格控制下进行的。计算机算法亦有相似的情况，其运行结果亦需研究者给出必要的分析判断。

计算机是出色的辅助工具，其算法可用以解决数字性问题，控制和指导仪器的运行，编辑处理文字信息，进行检索并寻找数据间内在的联系，建立数据库，等等。其中，后三项作用对生物信息学而言，至关重要。

三、不同类型计算机的功用

1. PC 机（Personal Computer）具有多方面的功能，可进行文字处理、表单分析、文件显示、互联网应用、甚至控制实验仪器。例如，登陆网站 http://www.axon.com 后，可以控制 AXON 的膜片钳(pCLAMP)。这是一种广泛应用于电生理学中，可以控制和测量神经元电活动的应用软件。还可进行离子成像和分析液浓度成像。PC 机在生物医学领域有多方面的用途，完成诸如：在基因组计划中可以分析 DNA 芯片的杂交信号、功能性神经外科手术中微电极的导向分析、诊断监测运动障碍（如帕金森氏症）等许多工作。PC 机和局域网（奔腾处理器、NT 工作站）的应用十分广泛，而且运算速度快、能力强大。这使得科学家无需使用超级计算机就能开展诸如构建分子结构和多序列对齐的工作。一般说来，机控的实验仪器有可选择的计算机界面，便可满足研究人员用于不同的实验目的。据估计，世界上仅有 1%的计算机微处理器用于台式 PC 机，其余的 99%则装备于其它的工业产品中，如飞机、供暖系统、实验装置、安全设施等等。它们通常是具有特定功能的软硬件结合的，且无需指令程序的芯片。

科学研究需要使用具有多种内置处理器的仪器，如气相色谱仪、电子天平、分光光度仪等等。例如，分光光度仪可在不同的波长处读取液体的吸光度，同时还可即时测量受测液体化学成份的变化。又如液相色谱仪，它的自动分离功能可根据分子的大小及溶解度的不同，将混合物分离成单一的组分。上述这类仪器均

由内置的微处理器直接控制，不必通过外接计算机另行控制。人们只需在类似 ATM （自动取款机）的小视窗上输入代码或提示指令，即可完成工作。近二十年来，这些微处理器从简单的控制回路发展成为今日可以贮存大量的数据和图象文件的芯片，正逐步取代记录信息的纸张和胶片的作用。

2. 超级计算机（Supercomputers）这类计算机可以完成需大量运算的任务，并且具有准确的工作记忆和超大容量的存储能力。它是网络服务商的主要的工作对象，大都应用 UNIX 操作系统，且已经服务于许多学术团体。鉴于其功能和价格，它所提供的服务已成为学术界设备共享的范例。例如，圣迭戈超级计算机中心(San Diego Super Computer Center，SDSC)(http://www.sdsc.edu/)是一家提供公共操作服务的公司，在学术界的应用十分广泛。SDSC 可提供并支持广泛的计算资源，目前其系列产品包括 CRAY C90 超级计算机、CRAY T3E 超级计算机、高级视像实验室、档案存储系统，等等。在美国，这些服务可供学术研究者及学生们使用。并且在费用均分的协议下，亦可供国内外的商业及政府人员使用。目前，已有超过 240 个机构的 5100 多位研究人员将这些服务平台用于科研中。

高速链接和并行运算

1997 年 6 月 20 日,匹兹堡超级计算机中心与德国斯图加特大学在大西洋两岸通过高速研究网络将超级计算机相互连接起来。这是首次将高速电信网用于跨洋运算。

作为国际高性能网络的预期模型，这项计划将匹兹堡的 512 位处理器 CRAY T3E 计算机与斯图加特高能计算中心的一台 512 位处理器 T3E 机相连。在不同的地点连接两台或更多的超级计算机，进行同一项运算任务称为宏观运算(Metacomputing)。这种连接实际上产生了 1024 位程序系统,其运算性能之高在理论上可达到每秒 6750 亿次。这项研究计划的实施是以两地间有高速越洋链接的一系列研究网络为基础的。近几年建立的这种网络，

其传输信息的速度比 INTERNET 快 100 倍。例如，vBNS（the very high speed Backbone Network Service）连接了美国国内多家大型计算机中心，其传输速度达每秒 6.22 亿个字节，一本完整的《大不列颠百科全书》可在 10 秒内传输完毕（摘自 PSC NEWS，http://www.PSC.edu）。

互联网是一个由转换器、路由器和光缆将计算机和工作站连接起来的网络系统。其强大的作用体现在它的交互模式上。现在已有了更高级的 PC 机系统，许多交互式任务在 WWW 上通过远程大型计算机完成。这就逐步取代了过去那种通过下载软件，再进行局域分析的方法。

四、计算机分析的局限性

英语字典、英文索引和其他许多的英语参考书，都是按照 26 个字母的顺序编排的。使用字典或其他参考书的方法，也是基于人们对这一套编排方案的理性认识上的。当然，语言有数百种之多，其他语言亦可以按其自身的字母顺序发挥各自的功用。一种语言是由很有限的一套符号组成的，是我们学习和认识事物的最佳体系之一。这就可以解释：为什么书籍能得到如此广泛的使用，并且成为一种稳定的信息交流和储藏形式。

与电脑中数字字符串的搜索相比较，识别字母这种方式可能会被简单地说成是“类比型识别方式”。人脑利用字母拼写来寻找某个电话号码，电脑则不是。后者更为快捷，电脑使用两种符号（0，1）来表示字母。由于搜索字符串查找的是整个文档，信息的分级对于以计算机为基础的查询方式并非必须。对比起来，人脑是用一种可视的方式去观察信息及其相互联系的，因此知道字母 X 排在字母 Y 之前。而电脑在没有结构性数据库的情况下，无效性和错误（即查询时没有结果）随着信息储存量的增大而增加。

如何教电脑像人那样按字母表顺序，去查找电话号码呢？当搜寻和分类整理信息（诸如： A 在 B 前，B 在 C 前等）时，我们需要把字母表的排列规则教给电脑。这是机器能够解释一系列分

级系统和优先权的命令语言。然而，我们需要的功能不只是简单的字符串搜索模式，而是电脑现今完成起来仍十分困难的一些工作（虽然对电脑来说，这些工作不需要太精确）。如果所要找寻的名字有拼写错误，电脑就不能查出。然而，当你手工查找一个名字时，即使对它的正确拼写并无把握，你也许仍能找到它。另外，你可能会找到你认为有意义的其他信息。虽然，你在开始工作前可能并没有期望会找到它。可见，电脑和人脑的工作方式是很不相同的。在科学研究中使用电脑查找信息，正如使用它查找电话号码一样。科学研究是一项人类探索发现未知世界的活动，探求物体间的相互关系是这一活动的中心工作。这就意味着，上述关系的定量化以及数字化后的解释是需要的。科学家必须对信息的字符串进行查询和对比，以确定字符串在配对过程中的质量。明白计算机工作的原理，以及计算机专家排列这些信息的方法，才能使计算机在科学实验工作中得到成功的应用（图 1-2）。

当一项实验结束时，所获得的大量数据需要进行分析处理。这包括建立各种数据库、进行统计学检验，等等。尤其重要的是，选出那些有分析价值的、能合理解释该实验的数据。这当然可以不借助于计算机。但这样一来，因常常依赖于实验者的经验，而不具有可靠性。比如，若实验在某一前提下进行，则期待的实验结果也以该前提为基础而产生了。因此，数据质量的评定往往应由旁观者得出。

科学家的直觉思维和主观想法是作出决策的重要因素之一。当然，计算机可以起到一定的作用，但只限于那些用于专业分析的软件，而且这一过程也得有人为干预。即使到了计算机可自动运行分析程序时，它也不能解释所有的数据。不过，用计算机辅助分析数据可逐渐消除实验人员的主观影响。因为计算机读取任何数据都一视同仁。

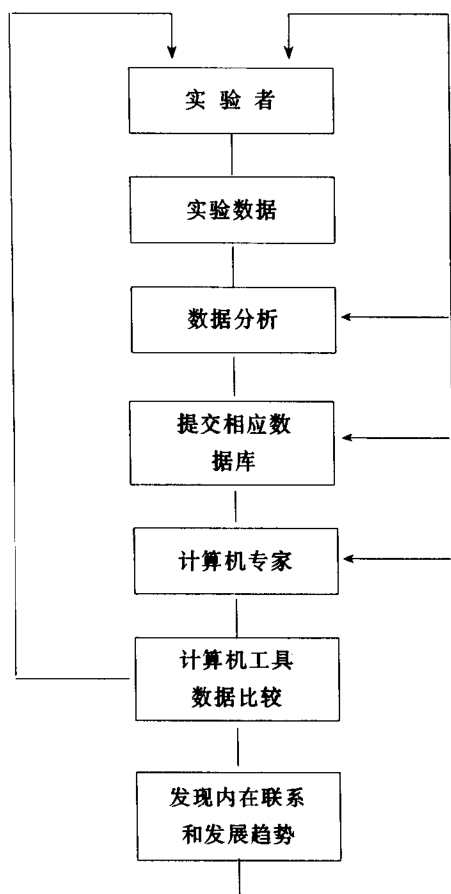


图 1-2　实验科学和计算机技术相结合

事实上，无论是以书面形式传播信息的实验技术手册，还是互联网，都不能取代实践的地位。在科研的过程中，最重要的是传授经验和获取信息。教科书、讲义和实验规章等书面上的知识并不能涵盖实践的全部内容。如今，互联网提供了一种很有帮助的服务——远程教育。现在有越来越多的教育机构网站（如：牛

津大学）就通过互联网提供了远程教育课程。互联网的“合作”特性可影响甚至误导人们的观念。实际上，互联网上的交互性（interactive）是有限的。

应当强调一点，机读数据的精确性依赖于操作人员的正确输入和操作。数据库相关注释的不可靠性、DNA和蛋白质库也有错误，这些事实已不是什么秘密。比如，从何处并如何得到这些序列的相关信息，就可能有错误。生物信息学的成功之处在于将全部注释可靠的数据库序列信息与准确的生物学数据直接联系起来。这种检验数据注释精确性的过程，是非自动化的，也需要花费许多时间。相反，为了确保数据库信息来源的可靠性，相关领域的许多专家们需逐字逐句地阅读这些信息。

研究者运用诸如BLAST这样的程序，对基因序列进行比较是相当地容易的，但要真正理解比较结果的含义却很困难。这些结果由比较序列的类型和起始序列而定。这二者均提到了一个事实，即除了核酸序列外，人们还需知道它们的功能结构和获取这些序列的细胞来源及其实验方法。换言之，需明白这些序列反映的生物学信息，才能理解它与相关序列或新的序列进行比较的结果。生命科学家在基因组计划的研究中发现大多数序列无生物学功能。因此，序列的来龙去脉对理解其生物学功能就显得尤为重要。研究人员希望能提出有预见性的信息，从而有利于实验设计，并用来迅速可靠的证实这些序列所反映的生物学意义。

五、对更好的计算机工具的需求

生物学和医学是多层面的学科。神经科学集中研究神经元的生物学特性；生物化学关注生物系统中的化学反应；分子生物学则是在分子水平研究生物间的相互作用，以及它们与作为整体的细胞和有机系统的相关性；病毒学和细菌学则关注着病毒和细菌各自的生物周期。生物学其他方面的研究多集中于特殊领域相关的特殊课题。生物学各方面的研究，既有差异又相互联系。随着数据管理技术的提高，这些领域间的交叉重叠会更加明显，因而

应用更高级的计算机工具已成为必需。

在过去的几十年中，生物学和医学均取得了非凡的进步。这就形成了一个正反馈环，即每一项新的研究成果都可成为促进这一领域发展和深化的推动力。因而，科研人员对一种高效且容量大的生物数据处理系统的需求目益迫切。但运行这一系统，若没有强有力的计算机支持是无法实现的。如今，计算机已成为生物学研究中必不可少的组成部分。没有了它，生物学和医学的发展毫无疑问地会受到阻碍。作为非生命体的计算机，在与生命学科（生物学）研究建立伙伴关系时，需要将两门学科中的某些部分进行合并。一些新的领域，如医学中的计算机应用和计算生物学，正逐渐兴起并迅速在生命科学中受到人们的重视。这些领域涉及到更为快速的生物数据分析功能，并逐步发现和拓展出许多以前未知的生物学发展趋势。这些趋势，在治疗方法的进步中发挥了作用（如药物设计等），达到延长生命、提高人们的生活质量的目的。

在日常应用中（如游戏、e－mail、文字处理及插图等),计算机一般都有人性化的界面。这些界面模拟了一个真实的桌面，堆放着大量的论文、文件夹、字典和回收站等。计算机荧屏界面上并没有软件和机器语言的编码，而是使用了一些有特殊含义的符号并最终把符号转换成电子线路中的一系列“开和关”的电流。

因为当今科学技术广泛地应用电脑并依赖电脑，所以类似的仪器已被电脑荧屏上相似的界面代替。这是一个必然的发展趋势，因为人用类比方式思考问题，而并非数字式的。我们需要看到一个图象，而并非一个数字表格。人们创立了三维图象法，以便了解那些从科学探索中获得的数字化的相互关系的含义。例如，我们可使用颜色来表示几乎全部的物理参数，如温度、电荷、密度、质量、高度和粘度等。

今天，在应用电脑网络工作的时代，象“虚拟细胞”这种词有了新的含义。在电脑网络系统的帮助下，可在细胞动力学中给定一个分级的相互关系，将已知的物理化学知识用数字来模型化。

这是一个需要超级计算机来完成三维物体（分子结构）复杂的图形、图象以及他们的动力（分子动力）的虚拟世界。

一个细胞中的原子数绝对是一个天文数字。在计算机模拟蛋白质功能结构的实验中，我们不仅要知道分子大小及相对位置等结构信息，还需知道分子中不同原子之间相互作用力的能量信息，这包括化学键、氢键、离子键等。我们不仅要了解一个蛋白质中的原子数目，更要知道这些原子之间的作用力的物理参数。计算的复杂性以及计算能力的不足（也就是数学算法的不足），使我们不得不注意有关生物分子的量子力学方面的描述。这些系统十分复杂，至今只有最简单的分子——氢分子的结构较为明确。经典力学和量子力学在一同探索着研究分子结构的捷径。

最小的氨基酸（甘氨酸）分子包括十个原子，小的蛋白包含 100 个氨基酸，估计每个蛋白质分子平均有几千个原子，而每个细胞又有几千个蛋白质分子。简单的推算，假设每个蛋白有 10^4 个原子，而每个细胞又含有 10^4 个蛋白 ——这相当于每个细胞中所有的蛋白包含了 1 亿个原子。最小的细菌基因组有 10 亿个原子。我们估计，包括所有的新陈代谢物质以及水分子在内，大约每个细菌细胞有 30 亿个原子。如果每个原子的位置由平均 5 个物理参数来决定，我们需要一个包含了 150 亿个空格的电子表格来储存这些信息。假如我们现在计算这个系统的动力，并想知道在 10 亿分之一秒后所有原子所处的位置。为了完成这个工作，需要 150 亿个计算步骤，用 100 MHz 的电脑需要 150 秒完成，计算 1 秒后的变化需要 150 万天，——也就是 4000 年以上。这样的分析过程无疑需要计算能力非常强的超级计算机。

六、网络与生物信息学

科学技术的发展不仅扩展了人们的视野，而且也改变了人们的行为方式。科学家本身的工作方法也应随着科学的发展及其自身的要求作出相应的改变。在传统上，科学家可以按个人的兴趣依靠自身的力量，在某一领域内进行研究。但今天，随着科学研

究的深入，他们愈发要以一种集体合作的模式来完成那些更复杂、难度更大的工作。在合作的过程中，信息的交流和共享是十分重要的一环。勿庸质疑，在网络时代，互联网已成为科学家们十分重要的辅助工具。

快速而便利的网上信息流对科学家来说有如下好处：首先，网络为任何一个研究者提供了获取信息的平等机会。充足的资料和自由的选取是决定科研的方向、起点的要素之一。其次，就分子生物学而言，生物学数据库中贮存了大量的信息。比如，已知的一些物种的某段碱基序列。虽然过多的此类信息显得富余，甚至过剩，但却为日后的综合比较分析提供了可能。而且，一旦需要应用某段信息时的便捷，则是不言而喻的。第三，有利于比较基因组学的形成和发展。这是一门跨种系的生物学新学科。现在认为，生物进化的研究显示出所有的生物体间有一种内在的联系，在本质上应是由同一原始的形式进化而来的。因此，比较不同种系基因组的异同是有实际意义的。对酵母等便于研究的低等生物进行比较的研究成果，就对我们进一步发现和理解人类的基因及其生化途径是有帮助的。

不同国家和地区的分子生物学研究者，应用不同的方法进行基因测序，其结果分别存放在不同的数据库中。但因为有了网络技术，这些揭示了相关基因的信息便有了自由传输交流的可能。研究者在确定一个新的基因时可能不必去筛选细胞系或动物组织，而是直接利用网上的公共数据库查询相关的目的基因。这就为人们的研究提供了一种新的方法，也可以说是一条捷径。

研究表明，基因的表达、蛋白质功能的发挥、生物信号的传导等生命现象，常常是多个基因、蛋白质相互作用的结果，机制复杂。对其深入地研究十分困难，而科学家们合作攻关、共享数据资源已成为日后科研的鲜明特色之一。另一方面，交流协作、信息畅通是生物信息学自身发展的基本条件。而网络在其中扮演了重要的角色。

将互联网引入生命科学的研究，其成果令人满意。它极大地增强了研究人员之间的各种交流，并减少了各领域中的重复研究。诸如美国国家生物技术信息中心（National Center for Biotechnology Information，NCBI）和欧洲生物信息研究所(Eutopean Bioinformatics Institute，EBI）等数据管理系统的出现，提高了世界上许多研究成果的完成效率，并将不同学科的生命科学家们有机地联合起来。生物学数据呈指数级扩增，这种扩增就需要专门的特定的处理系统生成有组织层次的数据清单，如区分蛋白质的生物学数据和多核苷酸（DNA 和 RNA)的生物学数据。蛋白质库(Protein Data Bank，PDB）是保存蛋白质数据的数据库。与大多数生物学服务机构一样，对于特定的蛋白质结构，PDB 也提供所贮存的分子间的相互联系及相互间的可能趋向。在 PDB 这些信息可从蛋白质的结构类型数据库（Structural Classification of Proteins，SCOP)来检索。象 SCOP 这样的数据库，是描述蛋白质大分子特征的有用工具。仅仅将这些分子分门别类地区分开来尚且不够，这种数据处理系统还应显示出一些能够引起研究人员兴趣的生物分子的相关信息。这些信息存在于一个特殊文件内，与其它相关站点的相关数据相链接。比如，小鼠肌红蛋白(myoglobin）的 PDB 文摘与肌红蛋白分子相关信息（如相关摘要）有多种链接可供选择，例如与其他生物的肌红蛋白分子相关链接（见图 1-3)。在给定一个数据查询项后，这个数据查询项的价值在于它可以与其它服务器上的相关记录链接，这样就可以提示出它与其它分子和其它系统的内在联系。

过去，在研究领域内一些无法触及的生物学问题，如今已是计算生物学研究的主题。生物化学、分子生物学、生物进化论、生物信息学、神经科学以及药理学仅仅是自然科学系统领域中的一部分，现已深受计算机应用的影响。与自然科学的其他学科相比，生物学领域至今仍有一些问题被认为是不可预测的，且生物学的许多方面互无关联。将计算机工具引入生命科学的研究领域，可

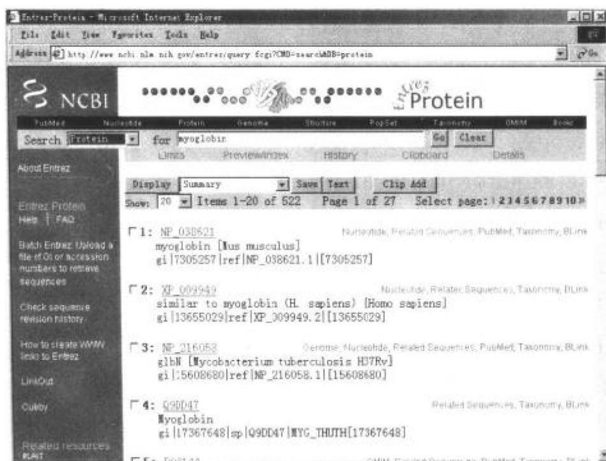


图 1-3　在 PDB 数据库中查询小鼠肌红蛋白（myoglobin）的结果，同时显示了多种其他生物肌红蛋白的相关链接。

大大地方便数据处理。更重要的是，应用计算机工具可以发现生物分子与各自相关领域之间的相互关系。生物学信息的极大丰富和预测能力的提高，使之明显地得到强化并且可很快地建立起各系统间的合作关系。将生物学视为可预测科学的设想，已成为许多研究人员的研究动力。没有这一点，生物学的科研目标将极大程度地受阻，甚至仅仅是人们的科学幻想而已。

在过去的几十年中，医学领域取得了许多重大的进展，生命科学也因此而倍受关注。一些有助于健康，提高人们生活质量的新药也投入了使用。这在很大程度上，促进了包括分子生物学和生物化学在内的多门生命学科的发展。生物信息学作为生命科学中的信息学科，将扮演重要的角色，影响并带动相关学科的进步与融合。

参考文献：

1. 刘秀艳，等. 应用计算机识别蛋白质功能。生命的化学，2000；

3 (20): 100-102

2. Fickett Jw. 通过计算机寻找基因。国外医学遗传学分册. 1998; 3, (21): 147-152
3. 范玉新,等. 基因功能分析与鉴定的新进展。国外医学分子生物学分册,1998; 3 (20): 100-103
4. 余才林,等. 基因组研究中寻找新基因的方法。国外医学遗传学分册, 1995; 4: 173-175
5. 李伟. 生物信息学新进展—第六届国际生物信息学和基因组研究年会综述。国外医学遗传学分册, 1999; 2 (22): 104-107
6. Altschul SF, et al. Basic Local alignment search tool. J Mol Biol, 1990, 215 (3): 403-410
7. Sussman JL, et al. Protein Data Bank (PDB): database of three dimensional Structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr, 1998, 54 (1): 1078-1084
8. Bartong J. SCOP: structural classfication of protein. Trends Biochem Sci, 1994, 19 (12): 554-555
9. Andreas D. Baxevanis, et al. 李衍达,等译. 生物信息学—基因和蛋白质分析的实用指南,清华大学出版社, 2000: 324-326
10. TIGR releases EST data publicly (news). Nat Biotechnol, 1997; 15 (5): 398

第二章　生物大分子

生物大分子主要是指蛋白质和核酸，它们存在于现在已知的所有的生命体中，是生命的标志。从化学组成上说，是生命与非生命的分界。在大千世界中，无论植物或动物，无论是高等或低等，这两类大分子物质，都是很相似的。蛋白质是由 20 种氨基酸组成的；所有的核酸是由数种基本的核苷酸组成的。生物大分子是与生命的基本现象，诸如新陈代谢、生长、运动、繁殖、遗传等等，密切相关的。因此，生命科学的研究必须从这两类大分子物质入手，对其要有深入的了解；用包括生物信息学在内的各种方法剖析它们的结构，并把结构与功能的研究联系起来。

生物信息学主要是随着分子生物学的发展而出现的一门的交叉科学，在人类基因组计划以及其它物种基因组计划中得以迅速的发展，相关的数据库及重要的理论、技术日益成熟完善。因此，其中的很多内容涉及到蛋白质及核酸的结构，并且生物信息学的研究方法以及相关软件和数据库的设计是以生物大分子的基本特性为基础的。这里，仅就相关知识作一简单的介绍。

第一节　蛋白质的结构与功能

蛋白质是细胞组织成分中含量最丰富、功能最多的高分子物质。它们在所有的生命体中起着关键性的作用。一个真核细胞中可以有多种蛋白质，它们的结构各异，亦各有其特殊功能。蛋白质功能的多样性是由其结构的千差万别所决定的。

一、蛋白质的结构

蛋白质是由氨基酸组成的，要理解蛋白质的基本特性需掌握

组成蛋白质的氨基酸的性质。

1. 氨基酸的结构

在自然界，氨基酸有 300 余种，但组成蛋白质分子的氨基酸仅有 20 种，且它们均属于 L-α 氨基酸（图 2-1）。

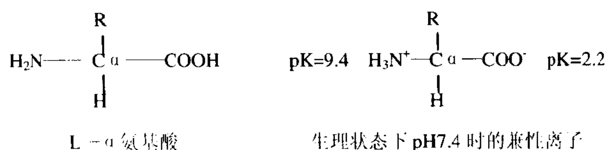


图 2-1 L-α 氨基酸的化学结构及其兼性离子特性：一个 L-α 氨基酸包含一个位于中间的 α 碳原子和 4 个化学取代基团，R 代表具有理化特性的侧链，NH_2 和 COOH 分别代表碱性氨基和酸性羧基。通常在生理状况下，二者均存在。

氨基酸有一个氨基和一个羧基，氨基增加其碱性，羧基与其酸性有关。在生理 pH 条件下，两端均带电荷，即氨基端质子化时，羧基端为去质子化状态。一种 α 氨基酸有别于其它 α 氨基酸的性质是由其残基或侧链决定的，通常认为这些残基是 R 基团。通过 R 基团的特征可区分 20 种 α 氨基酸中每一个的性质，其中，有些残基偏酸性，有些偏碱性，其它是中性的。

（1）酸性残基

谷氨酸根（E）和天门冬氨酸根（D）：

谷氨酸根和天门冬氨酸根分别是谷氨酸与天门冬氨酸的共轭化合物形式，生理 pH 条件下呈去质子化状态，带负电荷。

（2）碱性残基

赖氨酸（K）和精氨酸（R）：

赖氨酸和精氨酸在生理 pH 条件下呈碱性，呈质子化状态，带正电荷。

氨基酸的疏水性分类

氨基酸也可根据其疏水性分类。有些是疏水的，有些则是亲水的。根据残基的疏水性质，我们可以预计某些基因表达在蛋白质分子结构中所处的位置。疏水性残基往往位于蛋白质分子的核心处，而亲水性残基则主要分布于蛋白质分子的表面，与水环境相互作用。化学中“相似相溶”的概念在生物系统中同样适用，因此，在大多数生物系统中疏水—疏水的相互作用较相互对抗的疏水一亲水的相互作用更占优势。

· 谷氨酸、天冬氨酸、赖氨酸和精氨酸在生理 pH 条件下带电荷，主要位于蛋白质分子外部，与极性环境相互作用。一般来讲，带电分子亲极性环境，主要因为极性环境对电荷有稳定作用（比如：氢键、静电作用力等）。

· 丙氨酸、缬氨酸、亮氨酸、异亮氨酸、苯丙氨酸、蛋氨酸、甘氨酸、半胱氨酸和色氨酸是疏水性氨基酸，存在于蛋白质分子的核心以及其它疏水环境，残基中的碳链可增加其疏水性能。

· 天冬酰胺、谷氨酰胺、脯氨酸、丝氨酸和苏氨酸残基是不带电的极性基团，具有溶解的倾向。

· 酪氨酸的羧基增加亲水性，而其氨基侧链又具有疏水性，这种双重性使酪氨酸适于亲水、疏水两种环境。

· 组氨酸具有相对极性，其环型侧链结构的构象变化使其等电点范围较宽，并具有双重性质，组氨酸可根据环境的不同而呈质子化状态或去质子化状态。这一特性使其在许多酶的活性位点构成 Schiff 碱基。

表 2-1 为 α-氨基酸的一般性质，表 2-2 是被普遍引用的不同氨基酸的疏水性范围，其中的一些数值更为常用。但此表的重价值在于表中所列的不同的研究数值之间有着一致性。其中的正数数值代表着疏水性残基。

表 2-1　α-氨基酸的一般性质

氨基酸	单字母代号	疏水性	芳香性	脂肪酸	小残基	极性	电荷
丙氨酸	A	▲			▲		
精氨酸	R					▲	▲
天冬酰胺	N				▲	▲	
天冬氨酸	D				▲	▲	▲
半胱氨酸	C	▲			▲	▲	
谷氨酸	E					▲	▲
谷氨酰胺	Q					▲	
甘氨酸	G	▲			▲	▲	
组氨酸	H	▲	▲			▲	▲
异亮氨酸	I	▲		▲			
亮氨酸	L	▲		▲			
赖氨酸	K	▲				▲	▲
蛋氨酸	M	▲					
苯丙氨酸	F	▲	▲				
脯氨酸	P				▲		
丝氨酸	S				▲	▲	
苏氨酸	T				▲	▲	
色氨酸	W	▲	▲			▲	
酪氨酸	Y	▲	▲			▲	
缬氨酸	V	▲		▲			

▲　代表所列氨基酸残基特性或者部分特性

表 2-2　不同氨基酸的疏水性范围

残基	单字母代号	Kyte/Doolittle [1]	Edelman [2]	Eisenberg [3]
丙氨酸	A	1.8	0.4397	0.25
精氨酸	R	-4.5	-0.7010	-1.80
天冬酰胺	N	-3.5	-1.414	-0.64
天冬氨酸	D	-3.5	-2.588	-0.72
半胱氨酸	C	2.5	1.150	0.04
谷氨酸	E	-3.5	-1.270	-0.62
谷氨酰胺	Q	-3.5	-1.656	-0.69
甘氨酸	G	-0.4	-0.8634	0.16
组氨酸	H	-3.2	0.0268	-0.40
异亮氨酸	I	4.5	1.546	0.73
亮氨酸	L	3.8	1.517	0.53
赖氨酸	K	-3.9	-1.502	-1.10
蛋氨酸	M	1.9	1.746	0.26
苯丙氨酸	F	2.8	0.4345	0.61
脯氨酸	P	-1.6	-1.721	-0.07
丝氨酸	S	-0.8	-0.3841	-0.26
苏氨酸	T	-0.7	-0.0078	-0.18
色氨酸	W	-0.9	-0.0638	0.37
酪氨酸	Y	1.3	-0.4585	0.02
缬氨酸	V	4.2	0.5056	0.54

2. 肽键与肽链

(1) 肽键的构成

氨基酸之间借肽键相连，这基本上是由酸碱反应形成的。反应后失去一分子水（图 2-2)。为了进一步理解蛋白质骨架构象，需首先理解肽键的本质及其与多肽骨架构象的联系。肽键是一个特殊的键，约束着蛋白质的结构，形成一定角度，这种约束也构成了多肽的三维骨架结构。

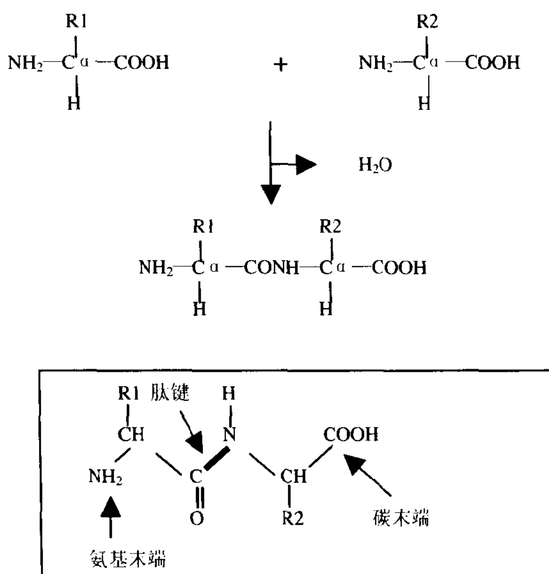


图 2-2 肽键的形成，两个氨基酸通过它们的氨基端和羧基端形成共价结合并释放一个水分子形成二肽，肽键形成一个平面称为酰胺平面，在 C-C 和 C-O 之间的共价键允许蛋白质折叠成复杂的三维结构。

(2) 肽键的性质

肽键具有双键性能（图 2-3)，因为形成肽键的原子间存在着

共轭系统。这种共轭性可稳定肽键的双键性能，使其具有刚性，不能以 C=N 为轴心旋转。肽键的旋转角称 Ω 角，在多肽链中很小，肽键的双键性能将参与肽键的原子约束于一个平面，使参数 Ω 角成 180。角。这部分是因为大多数二肽为顺式结构，在极罕见的情况下可出现分子构象转移， Ω 角变成 0°。这种分子构象变化多由脯氨酸残基引起。

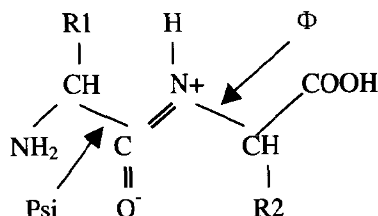


图 2-3　显示了肽键在一定程度上的双键性能，并标出 φ 角和 Psi 角。φ 角和 Psi 角是与多肽链骨架结构相关的主要成角构象，以二维形态描绘蛋白质骨架的三维构象，该区称为 Ramachandran 区。φ 角是 α碳原子和相邻的 N-H 基团的夹角，而 Psi 角是 α 碳原子与相邻 C 间的夹角。

（3）肽链

多肽链中肽键与 α-碳原子形成一条骨架，氨基酸的侧链在此骨架上向外伸出。多肽链中的氨基末端在左边，羧基末端放在肽链的右边。肽链中的氨基酸序列因蛋白质的不同而不同。每种蛋白质各有其固有的氨基酸序列。

3. 蛋白质的构象

具有生物功能的多肽及蛋白质都是有序结构，都有其一定的氨基酸百分比组成及排列顺序(一级结构),及特殊的高级结构(立体结构),即所谓的二级结构、三级结构及四级结构(并不是所有的蛋白质都有四级结构）。从而构成蛋白质折叠的途径（图 2-4）。

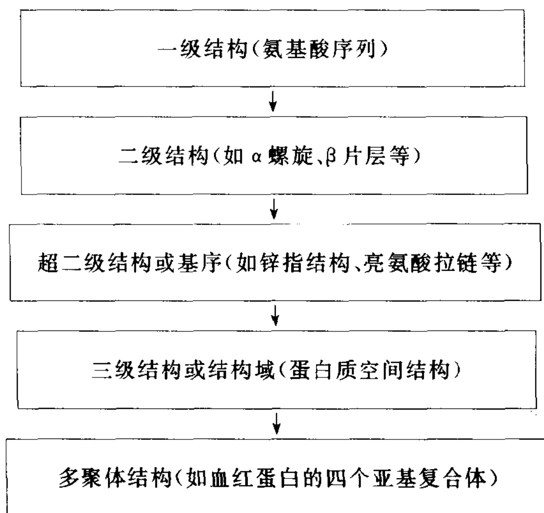


图 2-4　蛋白质折叠途径

蛋白质的一级结构是多肽链上氨基酸的线性排列顺序。了解蛋白质一级结构可更好地理解分子的三维结构。蛋白质的高级结构的形成主要依靠非共价键。每个非共价键的键能很小，但众多的非共价键足以提供很大的引力，以形成蛋白质的高级结构并保持其稳定。

以下是几种作用力，它们可使多肽链折叠并形成蛋白质的高级结构：

- 疏水键或疏水作用
- 静电引力
- 氢键
- 构象熵（Conformational entropy）
- 范德华力
- 共价键(如 二硫键)

(1)疏水作用与蛋白质折叠

疏水作用多在疏水性或非极性原子聚集处形成，远离水分子，

就象油分散在水中。由于油分子是疏水的，会自动聚集，可减少与极性水分子的接触。在室温下，这种作用主要是由熵驱动的。

疏水键一般被认为是蛋白质折叠过程的动力。大多数蛋白质所处的环境是在水中，由于水分子之间的氢键在不断地形成和断裂，蛋白分子的非极性侧链就形成这种聚集成簇的构象。利用其存在于水环境的条件，非极性基团埋于蛋白质分子内部，这有利于蛋白质分子的疏水区与水环境共存。

· 测定侧链疏水性的方法：

通常借分隔实验完成。Fauchere 和 Pliska 早期进行的实验得到了不同侧链的疏水性的评估值。他们将一种人工化合物投放入某介质中，代表蛋白质核心及其所处的水环境，再测定不同侧链的浓度。他们所用的介质是辛醇，它的长链脂肪酸和羧基未端与极性和非极性侧链均可发生作用，一些疏水范围和疏水值见表 2-2。

· ASA(accessible surface area，易接近的表面区域）及其与疏水性的关系：

易接近的表面区域（ASA）是水分子探针在溶质表面的作用位点(见图 2-5),ASA 可以增加可溶性分子的容量。侧链的非极性原子的 ASA 与其疏水性近似成线性关系。

(2)静电作用力

通常指蛋白质分子中离子对之间的静电引力。一般认为静电引力对蛋白质结构有特定的作用，这种作用力多遵循库伦定律。离子对所带电荷数，带电基团间的距离和相对介电常数是库仑定律的关键，这使蛋白质离子对与水环境之间的相互作用达到最大值。对所有带电离子对来说，主要是使其热力学的最适环境达到极限，独立带电基团存在于水表面，并被水分子完全溶解。

因此，将这些单个带电基团转移至蛋白质分子内部不大可能，有研究表明暴露于溶剂的带电离子对或单个基团可增强蛋白质分子结构的稳定性。

(3)氢键

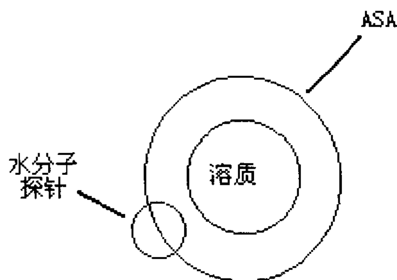


图 2-5 溶质易接近的表面区域。溶质在水中有一个特殊的表面称为范德华氏表面。水分子与溶质表面作用占据一定的空间，这个空间可以增加可溶性分子的容量。围绕范德华氏表面一圈的球形水分子中心附近区域决定水易接近的表面区域，这部分区域在模拟药物-受体相互作用中是非常有用的。

氢键对蛋白质结构及其稳定性起到多大的作用，并不十分明确。但氢键有规律地出现在α螺旋和β片层结构中,对蛋白质二级结构的形成可能有作用。侧链-侧链之间,侧链-主链之间也存在氢键。被埋在内部的氢键对蛋白质分子的结构有稳定作用，其主要原因可认为是蛋白质-溶剂、蛋白质-蛋白质氢键之间的竞争结果。

(4)范德华力

大多数蛋白质的折叠结构要利用分子中各种原子的紧密聚集，蛋白质分子的核心原子往往比表面的排列整齐。形成折叠结构时，蛋白质核心部位的原子往往较表面的或未折叠时的原子牢固。

(5)共价键

除了肽键以外，蛋白质结构中最有意义的共价键是二硫键。与不含二硫键的肽链相比，二硫键可使肽链受到一定程度地约束，从而起到稳固折叠结构的作用。一般来说，增加共价连接的长度，相应地会提高结构的稳定性。但这个规则仅适用于含有单个二疏键的肽链。在小分子蛋白质中，二硫键的稳固作用更大一些。

二、蛋白质功能

蛋白质是具有多种功能的生物大分子。

1. 酶的催化作用

酶是生物催化剂，绝大多数的酶是蛋白质。没有它，生命无法维持。酶的基本作用是加速生物化学反应，否则，就会因为反应速度过慢而无法维持生命活动。酶的催化底物具有特异性，其酶促反应效率与细胞内底物浓度密切相关。这种对底物浓度的依赖性避免了反应终产物生成过多，而最终使细胞自身受到损伤。酶对底物的专一性与酶的空间构象密切相关。尤其是酶的活性中心的空间构象，可使酶具有专一性，而与其它酶相区别。因此，了解酶的结构特征有助于我们掌握酶的功能。根据已知结构的蛋白质去理解蛋白质的结构特征，这对结构尚未明确的序列，在寻找其结构与功能的关系上，显得尤为重要。对那些重要蛋白质的结构与功能关系的深入了解，使人们掌握了控制这些酶活性的有力武器，并可有效地防止酶的失活。目前认为酶的失活与许多疾病的发生发展有关。

2.调节蛋白的作用

这些蛋白质的主要作用是调节细胞内的其它大分子的活性。调节过程与蛋白质的浓度有关，许多这类蛋白是通过负反馈调节机制发挥作用的。在大多数负反馈环路中，下游产物浓度增加会阻碍上游产物的形成。在 DNA 复制和 RNA 翻译中均可存在负反馈调节。

3. 存储

某些离子、代谢产物、或小分子可与蛋白质结合而保存在生物体内。例如，铁蛋白通过其亚铁血红素基团与铁离子的结合将铁离子贮存于肝脏内。

4.运输

有些蛋白质具有生物转运功能，转铁蛋白和血红蛋白就是两种转运蛋白，在体内分别运输铁离子和氧。

5.信号传递

有些蛋白质在生物体和细胞信号传递中有特殊的作用，它们大多是小分子和激素的细胞受体。结合小分子或激素后，产生信号

最终引发细胞内的反应。

6. 免疫作用

免疫系统中的大分子大多是蛋白质和多肽。例如，免疫球蛋白。这是一个庞大的蛋白质家族，参与多种免疫反应。

7. 形成细胞的结构

蛋白质中有相当一部分是结构蛋白，主要起机械支持作用。胶原蛋白是多细胞有机体内含量最多的结构蛋白，几乎所有组织都有胶原蛋白。

第二节　核酸的结构与功能

核酸是生命的遗传物质，是每一个已知生物体基因组的构成组分。这些分子由 DNA(脱氧核糖核酸)和 RNA(核糖核酸)构成，细胞用一些蛋白质作“工具”读取其基因组信息并翻译为其它蛋白质，以完成和控制细胞的活动过程，包括新陈代谢、产生生理信号、能量的贮存和转化以及细胞结构的构建等。

一、DNA 和 RNA 结构

核酸(nucleic acid)包括脱氧核糖核酸(deoxyribonucleic acid，DNA) 和核糖核酸(ribonnucleic acid,RNA)。前者是遗传信息的贮存和携带者；而后者则主要参与遗传信息的表达过程。

核酸亦称多核苷酸，是由数十个以至数千万的核苷酸构成的生物大分子。核苷酸是各种核酸的组成单位，它由碱基(base)、核糖(ribose)、脱氧核糖(deoxyribose)、以及磷酸(phosphate)几种分子连接而成。

在达尔文的物种起源学说发表了 10 年之后，当时还不知道核酸是遗传分子，但遗传物质的研究成为化学、医学、生物学的主要课题方向之一。1869 年，Friederich Miescher 首次从细胞中提取出一种物质，这种物质表现出酸的性质，溶解性依赖于溶液的 pH 值。他把这种物质命名为核酸，但却并不知晓这种物质正是基因的

实体结构。在 Gregor Mendel 证明种属的特异性可作为独立的实体结构遗传下去时"基因"这一构想就已产生了。DNA 作为基因的独特的载体，不仅已于 1944 年被 Oswald Avery 所证明，同时 Alfred Hershey 和 Margaret Chase 也分别证实了这一点。这距 Miescher 首次报道核酸的时间约为 75 年。九年后，James Watson 和 Sir Francis Crick 在 1953 年 4 月 5 日出版的一期英国的《自然》杂志上发表了一篇关于 DNA 分子双螺旋结构模型的论文。这一期杂志同时发表了 M. Wilkins 和 R. Franklin 关于 DNA 结晶的 X 线衍射研究的支持性论文。这一发现使得遗传密码在日后得以破译。这些密码决定了蛋白质生物合成所需的信息，并通过复制机制得以遗传，使子代具有与亲代基因组完全相同的拷贝。

核酸是由核苷酸组成的线性多聚体。核苷酸根据其芳香环碱基结构不同分为嘌呤和嘧啶两种——与磷酸核糖基团相连。它们的线性结构以三联体的形式按顺序读取。正是三联体中的核苷酸排列决定了每个基因的特异性及一种有机体区别于其它有机体的特征。RNA 是所有有机体中的主要核酸物质，由腺嘌呤核糖核苷(A)、鸟嘌呤核糖核苷(G)、胞嘧啶核糖核苷(C)和尿嘧啶核糖核苷(U)组成；DNA 是脱氧核糖核酸，四个碱基分别为腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和胸腺嘧啶(T)。DNA 携带遗传信息，其双链构象形成双螺旋结构，RNA 则是单链结构，但 RNA 可与 DNA 的单链杂交，在细胞内形成双螺旋结构。

双螺旋结构的稳定性由两股单链对应碱基对之间的静电作用力维持，并且 A 只能与 T 配对，G 也只能与 C 配对，这种精确配对，受核苷酸芳香环间氢键的热力学和构象限制。DNA 双螺旋是核酸二级结构的重要形式。

二、遗传密码

国外有学者认为：遗传密码的发现标志着生物信息学的开端。因为对每一种基因及其相应的蛋白质来讲，其 DNA、RNA 和氨基酸序列是唯一的。这一点，现在已经能够鉴定。科学家们很快发现，

早先的假设“一个基因对应一个蛋白”是不正确的，这是由于基因结构具有复杂的内含子和外显子组成，并且在所有生命形式中DNA 重组过程是非常丰富的。分子生物学家将这些过程在实验室中重复.获得成功 --产生了 DNA 重组技术- 这是基因工程的基础。

遗传密码的作用是将四种核苷酸的排列顺序翻译成 20 种氨基酸的排列顺序。每 3 个核苷酸为一组，决定蛋白质的一个氨基酸（图 2-6）。这种独特的过程在所有的细菌、动植物中均可观察到，并且翻译的顺序只能是从核酸到蛋白质，而不能反向进行。即核酸只能是蛋白质合成的模板，但蛋白质不能作为基因的模板，这就是分子生物学的中心法则，反映了所有生命有机体生长繁殖的过程。

1. The Standard Code (transl_table=1)

By default all transl_table in GenBank flatfiles are equal to id 1, and this is not shown. When transl_table is not equal to id 1, it is shown as a qualifier on the CDS feature

TTT	F Phe	TCT	S Ser	TAT	Y Tyr	TGT	C Cys
TTC	F Phe	TCC	S Ser	TAC	Y Tyr	TGC	C Cys
TTA	L Leu	TCA	S Ser	TAA	* Ter	TGA	* Ter
TTG	L Leu i	TCG	S Ser	TAG	* Ter	TGG	W Trp
CTT	L Leu	CCT	P Pro	CAT	H His	CGT	R Arg
CTC	L Leu	CCC	P Pro	CAC	H His	CGC	R Arg
CTA	L Leu	CCA	P Pro	CAA	Q Gln	CGA	R Arg
CTG	L Leu i	CCG	P Pro	CAG	Q Gln	CGG	R Arg
ATT	I Ile	ACT	T Thr	AAT	N Asn	AGT	S Ser
ATC	I Ile	ACC	T Thr	AAC	N Asn	AGC	S Ser
ATA	I Ile	ACA	T Thr	AAA	K Lys	AGA	R Arg
ATG	M Met i	ACG	T Thr	AAG	K Lys	AGG	R Arg
GTT	V Val	GCT	A Ala	GAT	D Asp	GGT	G Gly
GTC	V Val	GCC	A Ala	GAC	D Asp	GGC	G Gly
GTA	V Val	GCA	A Ala	GAA	E Glu	GGA	G Gly
GTG	V Val	GCG	A Ala	GAG	E Glu	GGG	G Gly

Click here to change format

图 2-6 NCBI 物种分类站点的标准遗传密码表，该站点还列出广许多种生物的遗传密码使用表以及与标准遗传密码不同之处。

一些病毒的基因组中只有 RNA 分子,根据中心法则,人们预计这些病毒直接以 RNA 在宿主细胞内指导蛋白质的合成。而实际上，进入宿主细胞后，病毒依靠一种反转录蛋白将 RNA 转录为

DNA，称为互补 DNA 或 cDNA。再由一种称为整和酶的蛋白催化，将 cDNA 插入宿主基因组中。以 RNA 为模板合成 DNA 的过程，在非病毒生物体内并不发生。催化这一过程的酶称逆转录酶，逆转录酶在实验室的应用使分子生物学取得了巨大的进步。

逆转录酶以及嗜热菌体内的热稳定 DNA 聚合酶是当代分子生物学重要的研究工具。应用聚合酶链式反应（PCR）和 cDNA 合成技术可以迅速地从一小片组织中鉴别出新的基因。在过去，这些方法多用于法医学的研究和实践中。如今，在应用此技术的基础上，新的 DNA 序列（以及氨基酸序列）不断地被发现。为了存贮、处理以及破译这些遗传信息，生物信息学在其中发挥着独特的作用。

遗传密码具有通用性，在数量上多于其编码的氨基酸的数目。明白这一点对理解生物体基因与其蛋白质结构和生命多样性之间的关系至关重要。

遗传密码的数目是冗余的，三联体密码（密码子）有 64 种，而编码的氨基酸仅有 20 种。一些氨基酸由一种以上密码子编码，包括终止码。其中起始码也编码蛋氨酸。这说明氨基酸序列比 DNA 序列更保守。这与进化的机制有密切的联系。因为有些 DNA 的点突变并不引起氨基酸序列的变化，这种点突变称为‘沉默突变”。由于它们对表现型无影响，因而不受自然选择的作用。有机体并不使用全部的密码子，而往往只选择一种，从而使密码子过剩现象受到限制。这称为密码子惯用性或倾向性。这种选择性在不同生物体是不一样的，密码子倾向性在应用 DNA 重组技术时非常重要。当一种生物的基因克隆后重组到另一生物的基因组内时，有的情况下密码子的倾向性可导致合成无功能的蛋白质，或者影响蛋白质合成的水平。密码子倾向性也可作为一种保护机制，对抗外源性病原体 DNA 的侵袭。机体未使用的密码子对其它外源 DNA 可起到终止码的作用，从而有效的抑制了病原体繁殖所必须的功能蛋白质的合成。

密码子倾向性是遗传密码通用性的结果。除一些细胞器的DNA外，所有生物体，包括病毒在内，都使用相同的密码子编码20种氨基酸用于蛋白质合成。这意味着基因可在不同生物体间传递，这也是现代生物技术工程的基础，从而也使生物信息学的研究不必依据细胞起源来辨别基因序列所表达的信息。这种通用性，增加了数据库中进行统计分析的DNA序列的样本含量。同时，使动物模型如小鼠、果蝇、大肠杆菌的基因(这是我们所熟知的)与人类相应基因进行比较变得更为容易。DNA序列的相似性可使相关生物体的基因得以快速鉴定、克隆和测序。另外，一种生物体丢失的生物信息也可根据这一点来推测。据此，我们也可将果蝇和线虫等模式生物体内的“工作过程”与人的新陈代谢进行类比研究。

三、基因与进化

DNA是生物体的遗传物质。在80多年前，有人开始用gene一词来表示遗传物质。那时，对基因与DNA之间的关系几乎一无所知，因此有人认为基因是没有物质基础的空洞无物的概念。甚至时至上个世纪50年代，有教科书还坚持认为基因不过是一种唯心的臆测。现在，人们清楚地认识到：基因是DNA大分子上的一个个片段，有复制、转录等主要功能，是生物遗传繁殖的物质基础。

基因是所有生命的遗传单位，由脱氧核糖核酸即DNA组成，但有些病毒如人类免疫缺陷病毒(HIV)，为反转录病毒，其基因组由RNA构成。根据基因组的形态，即有无细胞核出现，所有的生物体可分为两大类，真核生物和原核生物。原核生物是单细胞生物，又可分为两个王国：真细菌属和原生质属。

尽管形态特征上相似，但真细菌与原生质的基因组结构是不同的。三大生物王国的分型(指真核生物、真细菌和原生质)是由分析各自的核糖体RNA而得出的。但是根据许多蛋白质和新陈代谢途径的研究结果，原生质属分支也表现出真核生物和真细菌某些特征。

生物信息学为解决生物进化树的正确顺序和正确分类问题提

供了一种新的分析工具。很明显，仅靠基因型分类或表现型分类是无法解决这一问题的。分子生物学并不能简单地取代进化生物学较早的分支学科(诸如从形态、结构来进行生物的分类研究),它更是一种补充。针对两种基本基因组结构的起源和共存，对于谁先产生和二者如何共存问题，进化生物学家们展开了激烈的讨论。包含了三大生物王国内全部生物的基因组计划的开展，将有助于我们对这一问题解决。比较基因组学从分子的水平来研究进化的现象，它有利于我们了解缤纷繁复的生物之间的共性和本质。

· 真细菌

真细菌是单细胞原核生物，具有高度聚集基因结构和组织的基因组。所有基因都含一个编码区，编码相应蛋白质的氨基酸序列；与编码区相连的是控制区，它决定 DNA 转录和翻译时蛋白质结合 DNA 的方式。通常，基因被分成一个个功能单元，被和谐地调节着。这些基因可编码出构成新陈代谢通路的多种酶，这种相应的多基因结构称操纵子。操纵子反映的是基因的功能单元，及其上调和下调作用，比如，基因表达，由单一的转录单元协同完成。

• 真核生物和原生质

原生质和真核生物的基因组结构较真细菌复杂，它们的基因不再是简易、单一的编码框,而是由外显子(编码区)和内含子(非编码区）共同构成的。真核生物基因组含 5%～15%的编码区,即基因,有大量的 DNA 并不编码蛋白质，而且至今人们不知道其功能。这些 DNA 可能在基因组中具有重要的调节作用，尤其是在减数分裂中发挥作用。减数分裂是有性生殖中染色体 DNA 重组的重要过程。

原生质属在形态上与真细菌很难区别，在基因组结构和部分新陈代谢途径上与真核生物更接近。在一些已完成测序的原生质属基因组中，这一点更为明显。但是，生物界的分类主要依据参与代谢的酶的类别而定的。原生质属可能是最古老的生物，或者是现存生物中与推测出的地球生命的共同祖先最为相似的生物种属。

同样，遗传密码的普遍存在暗示了单细胞是所有生物的共同祖先

参考文献：

1. Kyte J, Doolittle RF. A simple method for displaying the hydropthic character of a protein. J Mol Biol, 1982, 157(1): 105-132
2. Edelman J. Quadratic minimization of predictors for protein secondary structure. Application to transmembrane alpha-helices. J Mol Biol, 1993, 232(1):165-191
3. Eisenberg J, et al. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol, 1984, 179(1):125-142
4. Fauchere JL, et al. Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res, 1988, 32(4):269-278
5. Richards FM, Richmond T. Solvents interfaces and protein structure. Ciba Found Symp, 1977, 60:23-45
6. Chothia C. Hydrophobic bonding and accessible surface area in protein. Nature, 1974, 248(446):338-339
7. Miller S, et al. The accessible surface area and stability of oligomeric protein. Nature, 1987, 328(6133):834-836
8. Serrano L, et al. The folding of an enzyme. II. J Mol Biol, 1992, 224(3):783-804
9. Baker EN, Hubbard RE. Hydrogen bonding in globular protein. Prog Biophys Mol Biol, 1984, 44(2):97-179
10. 顾天爵等,生物化学(第四版)。人民卫生出版社,1997,P6-14

第三章　数据库和搜索工具

今天，生物信息学已成为生命科学最为活跃的研究领域之一。数据库是生物信息学重要的工作平台，是其基本构成之一。各种各样的生物学数据库不断出现，其数量增长十分迅速，同时数据库的内部结构亦日趋复杂。Nucleic Acid Research 杂志每年第一期都公布互联网上最新的生物学数据库资源，2001 年 1 月公布的数据库有 280 个，而这个数字在 2002 年 1 月增长到了 349 个。且数据库的类型更加丰富，专业性更强，几乎覆盖了生命科学的各个领域。目前，这类数据库的服务已实现了高度的计算机和网络化。算法和软件的进步、数据库的一体化、服务器-客户模式的建立使之成为生物、医药、农业等学科的强有力的研究工具。因人类基因组等各类计划的实施，也促使数据库中的数据以极高的速度增长。至 2002 年 1 月，GenBank 已有 1,458.5 万条核酸序列，SWISS-PROT 有 10,2387 条蛋白质序列，PDB 收录 17,082 套结构信息。

数据库涉及到的内容主要包括两大部分：数据库组织和数据库开发工具。前者有著名的美国国家生物技术信息中心（NCBI）、欧洲生物信息学研究所（EBI）以及日本生物信息学服务器（GenomeNet-Japan）等等；而后者则包括同源序列搜索基本工具 BLAST 和 FASTA（Search Tools）等。本书将就有关内容作一介绍 。

第一节　计算机工具和数据库

当前，生命科学家正致力于寻找所有生命体中内在的遗传编码，及其在对抗病原体时的意义。阐明 DNA 分子的三联密码子及

其与翻译产物的关系是了解绝大部分以碳元素为基础的内在生命形式、机制的第一步。生物学不仅仅是一门关于生命体实验性研究的学科，而且在过去几十年中，生物数据的指数性增长已给这门学问增加了预测性的内容。今天，人们运用基础物理和化学的相关定律，使得许多生物学现象得以解释。利用不断丰富的数据，许多现存的生物学疑团将最终得到解决。这将促进新的生物规律及其定律的发现，并且有利于我们对那些十分复杂的生命体系有一个深入的理解。

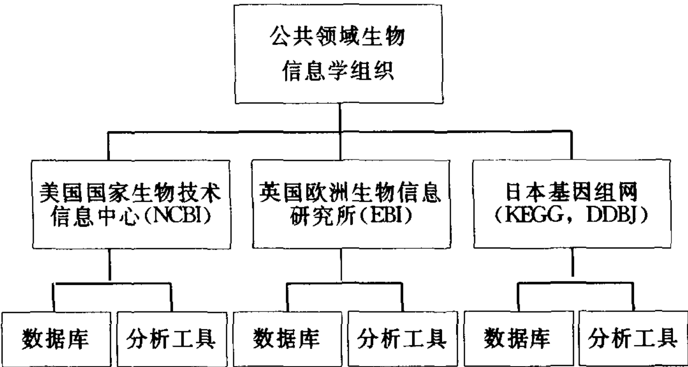


图 3-1　公共领域生物信息学服务器

管理、识别和鉴定那些指数级增长的生物数据时，计算机工具和数据库最为重要（图 3-1)。美国国家生物技术信息中心（NCBI)和欧洲生物信息研究所(EBI 是两家主要的生命科学信息服务机构，负责处理这些十分庞大的数据。它们拥有的可靠的数据库和分析软件，是当前生命科学界极具价值的研究工具。每天有大量新提交的条目进入它们的数据库，这些机构的职员将新的数据添加到适当的数据库中。这将保证那些订阅他们数据库的科学界同仁能及时地更新知识，促进了各门学科的进步。而这些机构（比如 NCBI 和 EBI)提供的服务又是由快捷、容量大的计算机实现的，这

些计算机能够完成必要的分析任务；因特网界面则使这种电子对话更易于进行。

一、美国国家生物技术信息中心(NCBI)

1988 年 11 月，美国参议院意识到构建计算机数据处理系统在生物医学和生物化学领域中的必要性，并通过法案帮助国家医学图书馆(National Library of Medicine，NLM)建立国家生物技术信息中心（National Center for Biotechnology Information，NCBI）图 3-2)。国家医学图书馆的任务主要是维护生物医学数据，而国家生物技术信息中心则特别涉足新型分析软件的开发，以帮助理解在致病过程中扮演关键角色的分子和遗传的活动过程。其主要的四大任务是：

(1) 创建适合分析、储存分子生物学、遗传学及生物化学等各类数据的自动化仪器；

(2) 促进科学界(如科研人员、医学工作者等)对可获得的数据库和分析软件的使用；

(3)整理全球的科研成果，收集生物学数据；

(4）对于重要的生物分子的结构-功能关系实施计算机分析、研究。

国家生物技术信息中心分为三个分支机构：计算生物学分部（Computational Biology Branch）、信息工程分部（Information Engineering Branch）和信息资源分部（Information Resources Branch)。NCBI 的科研人员包括计算机专家、分子生物学家、数学家、生化学家、医学研究人员和结构生物学家。国家生物技术信息中心工作人员通力合作，运用数学和计算机工具研究多种疾病的分子基础。其研究的三个主要层面是：

(1)对人们关注的基因及基因产物的序列进行分析；

(2)更好地理解、分析基因的组成；

(3)预测被研究分子的结构(如蛋白质)。

分析步骤包括：通过将新的蛋白或多核苷酸基因序列与已知

基因序列或蛋白序列进行比较，从而得到与新序列同源的已知基因或蛋白。对于那些功能尚不清楚或在数据库中缺乏已知同源基因的新基因而言，了解该基因的整个基因组构成可能是很有效的方法。用结构已知的同源基因预测结构未知的分子结构，使我们能预测结构及功能尚不清楚的分子可能具有的功能特性。

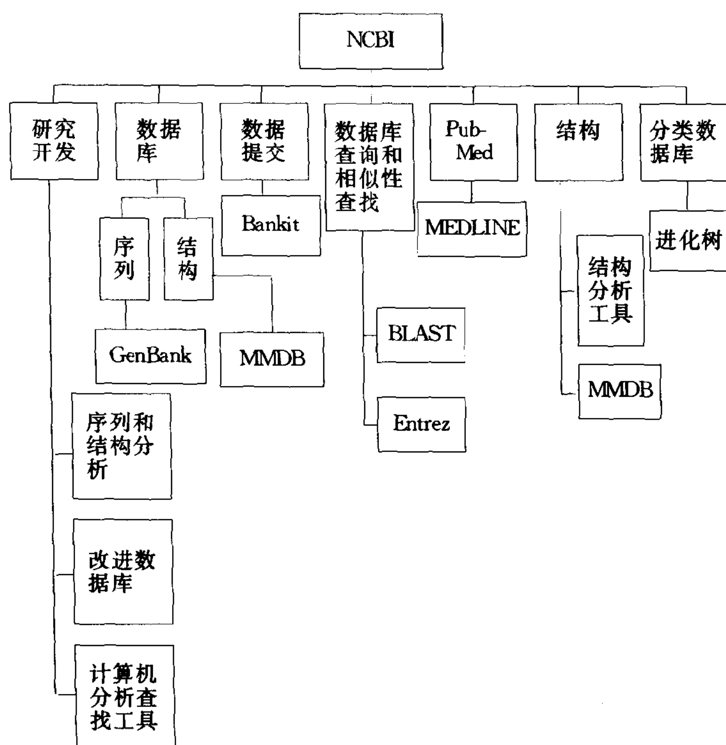


图 3-2　美国国家生物技术信息中心（NCBI)结构图

国家生物技术信息中心(NCBI)支持的数据库：

1.蛋白质序列，即实验获得的蛋白质序列以及来源于核苷酸文库的核苷酸翻译序列。

(1) 冗余蛋白质序列库（如 PIR 数据库，由 PIR1＋PIR2＋PIR3 组成）

(2) 非冗余或冗余较少的蛋白质序列库（如 NR、SWISS-PROT、PDB）

2. 核苷酸序列(DNA 和 RNA)，这些 DNA 和 RNA 序列来源于非自动测序计划（如： GenBank），或自动测序仪（如：dbEST）。

(1)冗余核苷酸序列数据库（如 dbEST）

(2) 非冗余或冗余较少的核苷酸序列数据库（如 GenBank）

冗余性的含义

由于许多科学工作者的研究是各自进行的，这就导致了在确定基因和蛋白质时使用了重复的命名。这个问题在新的研究领域提交数据的过程中，显得非常突出。这样，同一 DNA 序列在不同名称、记录和注释下，出现了不止一次。而只有这一领域的专家才能认出，这些看似不同的记录其实指的是同一事物。这就好象在电话号码本上以三种形式——姓、名、昵称来写你的名字，而每一个都是指你本人，其后的电话号码是相同的。冗余性在生物学数据库中是一个相当复杂的问题。每个数据库都有各自的冗余序列定义。大多数数据库，尤其是在大型的研究计划中，多采用自动方式解释冗余性。这种方法没有手工干预定量准确，但速度快。另一方面，为了数据的完整性，非冗余数据库也允许有冗余信息序列。当然，冗余性也有用武之地，如在数据库处理 DNA 序列时。两个相互竞争的实验室，有时可能会发表同一基因的核酸序列，但却有一个或几个碱基不同的情况。这是因为选择不同株别的研究动物而存在突变所造成的差异，还是测序上的错误呢？如果这个基因均取自同一生物体，那么说明有一方在测序上存在错误。事实上，这是一个无意识的质量控制的过程。

NCBI 中使用频率最高的蛋白序列数据库清单：

• Alu 这是一套精选出来的已翻译的 alu 重复序列，在查询序列中，它可以标出可能的 alu 重复序列。这一数据库还可通过匿

名FTP 在 NCBI 查询。(在／pub/jmc/alu 目录下)

· E. coli 这是只负责装载 E. coli 基因组 CDS 的数据库。

· Kabat：有关免疫学的序列数据库。

· Month 这是一个展示新近(最近 30 天以内) 修订资料的数据库，包含从 GenBank 获得的 CDS 翻译序列以及从 PDB、SWISS-PROT 和 PIR 数据库获得的其他的蛋白序列提交条目。

· NR(non-redunant)这是一个包括所有非冗余 CDS 翻译序列数据库,包含从 GenBank 获得的 CDS 翻译序列以及从 PDB、SWISS-PROT 和 PIR 数据库中获得的蛋白序列存储信息。在此数据库中，相同序列的蛋白合并为单一存储信息。

· PDB：包含三维结构已知的蛋白序列。蛋白质数据库(Protein Data Bank,PDB)组织位于纽约长岛的 Brookhaven 国家实验室。这些信息还可从 MMDB——NCBI 的 PDB 镜相站点获得。PDB 的绝大多数条目是非冗余的，如果同一序列条目有多个结构信息，就保留质量高的那一个。如对于晶体结构，采用其分辨值最小(如 1.8 埃优于 2.2 埃)的条目。然而,其他变量,如结合了金属或配体的复杂结构，在特定的生物中同一序列允许有多种结构信息存在。

· SWISS-PROT：这是一个来源于 SWISS-PROT 数据库的最近发布的蛋白序列条目的数据库。现由 EMBL(欧州分子生物学实验室)的分部 EBI 支持。这是一个通过信息相互参照获得的蛋白质序列数据库，可以从互联网上获得。 SWISS-PROT 是非冗余数据库，由日内瓦大学的 Amos Bairoch 维护。

· Yeast 酵母(S. cerevisinae)蛋白序列数据库,储存从酵母蛋白测序计划中获得的序列。

NCBI 应用频率最多的核苷酸数据库清单：

· alu：它允许在查询序列中标出可能的 alu 重复序列。这一数据库还可通过匿名 FTP 经 NCBI 检索。(在／pub/jmc/alu 目录下)

• dbEST 这是个收录 GenBank、EMBL、DDBJ 中 EST 条目的非冗余数据库。ESTs 是通过自动测序得到的 cDNA 一端的序列,很少有人为的干预因素。因此,相对于其他序列数据库,这将增加错误率。最常见的错误为测序错误、异源序列污染(heterologous sequence contaminations)和转录重复序列。

• dbSTS 这是一个收集 GenBank、EMBL、DDBJ 的 STS 条目的非冗余数据库。

• E. coli 只包含 E. coli 基因组核苷酸序列。

• EPD：真核细胞启动子数据库（eukaryotic promotor database），包含公共数据库中已知的所有真核细胞启动子序列。

• GSS：基因组纵览序列(Genome Survey sequence),基因组序列数据的一端数据、外显子捕获序列及 alu PCR 序列。

• HTGS：高通量基因组序列数据库（high throughput genomic sequence).

• Kabat：处理免疫学问题的序列数据库。

• Mito 专门处理线粒体序列的数据库。

• Month这是一个展示新近(最近 30 天以内)修订条目的数据库,其条目可以在 GenBank、EMBL、DDBJ 和 PDB 序列数据库中同时查到。

• nr 这是一个非冗余的 GenBank、EMBL、DDBJ 和 PDB 全部序列的数据库。它不包含 EST、STS、GSS 或 HTGS 的序列条目。完全相同序列条目只出现一次。

• PDB：这个数据库的序列来源于已知三维结构的分子。

• Vector: 这是 GenBank 的载体序列子集库。

• Yeast: 酵母 (S. cerevisiae)基因组核苷酸序列数据库,储存从酵母基因组计划及其他相关的酵母测序计划中获得的序列。

NCBI 提供的主要服务：

NCBI 网站的服务器提供以下 8 种主要数据库和分析工具：

(1) PubMed（Public MEDLINE）

(2) BLAST(Basic Local Aligment Search Tool)

(3) Entrez

(4) BankIt

(5) OMIM(Online Mendelian Inheritance in Man)

(6) Taxonomy

(7) Structure

(8) Books

1. PubMed

PubMed 是国家医学图书馆(NLM)的搜索服务器,用户可以在 MEDLINE 和 pre-MEDLINE 上获得超过 11,000,000 个引用条文(截至 2002 年 1月)。与网上杂志及相关数据库的链接,可以使用户方便快捷地检索到有关信息,而且目前可以通过相关链接得到许多种杂志的全文。它可以用关键词检索含有相关主题的期刊文章。为了增加查找的针对性,还可应用多个关键词。为方便用户,作者姓名和杂志名称也可作为查询标准。

2. BLAST:(Basic Local Alignment Search Tool)

局部排列基本搜索工具(Basic Local Alignment Search Tool)是一套相似性搜寻程序,可识别特定序列的分类和可能的同源性。这些程序的功能强大,能分析 DNA 和蛋白序列。其详细情况将在有关的章节里给予介绍。

3. Entrez

研究人员有义务编制出原始的非冗余的数据资料,以方便人们对特定规律的理解。为避免或减少发表材料中的冗余现象,研究人员必须确保自身工作的原创性,这并非易事。但相关数据库中详尽的查询工具可使这件工作变得相对容易一些。例如,如果研究人员拟确认某蛋白质家族的一个特性,很明显其下一步工作是确保该项研究具有新意,换句话说,这是新发现吗?为回答这一不可小视的问题,必须查询包含了相关关键词的所有引用条文。于是,可能会有如下三种结果:第一,你进入了充满冗余信息的死胡同。在

这种情况下，所查资料与以前的研究相同或非常接近。这时，聪明的研究者就会停止对冗余数据的研究，而转向关注其他方面的课题。第二种情况是没有相似的发现，此时该研究与已知的引用条文完全无关。这是好事，也可能是坏事。它可能是重要的原始发现，但也可能是各种谬误所为。此时，研究者必须进一步加以调查、验证原始记录，以确保其发现的正确性。或者，找出原始记录以及自己工作中的潜在错误。第三，查询到的相关信息，既是对先行研究的支持，又不是以往研究的重复。对于研究者来说，这是一种理想的情况。相关的引证可以作为此后深入研究的支持性参考。无论如何，进行可靠而有创新性的研究，研究者必须运用搜索引擎。它不仅高效而且可以使人们获得定期更新的相关数据资料。一般来说，这些由政府支持的搜索工具是现有公共领域内可信度最高的软件，且具有友好的界面。研究人员可以很容易地从互联网上获取。

NCBI 下的 Entrez 是最受欢迎的搜索引擎之一。Entrez 网页界面(http://www.ncbi.nlm.nih.gov/Entrez/）允许用户从众多可靠的数据库中获得文献目录和生物学资料。例如，可从 SWISS-PROT、PDB、PIR 以及 PRF 检索蛋白序列信息。从 Brookhaven PDB 检索结构已知的蛋白信息，这些蛋白已并入 NCBI 的分子模型数据库，也叫做 MMDB。翻译的蛋白和 DNA 序列可从其上一级 DNA 序列数据库（如：GenBank、EMBL 和 DDBJ）中检索。就文献目录或引证搜索而言， Entrez 利用 PubMed 的文献目录数据库，可在 MEDLINE 和 pre-MEDLINE 上获得超过 11,000,000 篇生物医学文章。通过 Entrez 也可获得染色体基因定位和基因组的数据。对某一特定的搜索，Entrez 提供多个标准。例如，搜索某一相应数据库，你可以发现从一个给定的单词开始的所有可能术语。在术语后加一星号，Entrez 可搜索到所有以该术语开头的词条。如搜索“ inter * ”可得到以 inter 开头的所有术语，如 interstetium、intermolecular 等等。还可以利用 Entrez 的智能搜

索，即利用短语或多组单词搜索。Entrez 可以将相关术语组织到一起而排除掉不相关的术语。

例如，为了查找某一指定作者（如：Wu M）关于某一给定主题（如：apoptosis）的所有可能引文，用户可以键入有关作者的术语（如：Wu M）及感兴趣的主题·(如：apoptosis)。Entrez将自动识别和组织相关的术语（如：作者的姓及字母的大小写），使搜索引擎从有关 Wu M 和 apoptosis 中寻找到所有的相关资料（“Wu M ”及 apoptosis）。使用自动组词功能，Entrez还可以组词。否则，将被视为分开的术语。插入引号，可使 Entrez 将似乎不相关的多个术语视为一个（如“ brca 1”)。然而，NCBI 建议用户只让 Entrez 组织一些特定的术语以减少不准确的检索。如果检索的清单太长，Entrez将终止搜索操作，并且会提醒用户。

通过标志符检索是查询某一特定引文或序列最精确的方法之一。标志符是一种索引数字，在相关数据库中为特定的序列或文章指定的一个标志符号。例如，MEDLINE 引文标志符指的是 UID 码，而属于序列的标志符称做 GI 码。检索 MEDLINE 的 UID88067898 引文，用户只需在 Entrez 搜索引擎输入 UID88067898，就能找出这一被指定 UID 的MEDLINE 引文。

Entrez 上有大量的搜索项目。由于有适应性的属性，有经验的用户会发现它们非常有用并能节约时间。以下介绍的一些搜索项目，可以满足用户的特殊需要：

- Keyword 允许用户搜索一套专门的指定术语。这些术语与 NCBI 可使用的数据库相关（如 GenBank、EMBL、PDB、DDBJ、SWISS-PROT、PIR 或 PRF)。

- Accession 允许用户搜索与蛋白、核苷酸序列、结构或者基因组的记录相对应的序列号。

- Author Name 含有发表论文的作者的相关信息资料。这是 MEDLINE特有的项目。

- Affiliation 用于搜索作者的所属单位和地址。

• Journal Title 用于搜索发表文章所在的期刊名称。用户可以应用 List Terms，浏览期刊缩写名清单（如 the Journal of Biological Chemistry 缩写成 J Biol Chem），以便查找。

• E. C. Number 是由酶学委员会分配给各种酶的指定码。

• Feature Key 是用于搜索表示某种 DNA 特定属性的关键词的一个搜索项目。

• Gene Symbol 用来搜索给定基因的标准名称。

• MEDLINE UID 用 MEDLINE 标志符搜索引文。

• MeSH Terms 用于搜索 MeSH 主题词。是一套为 MEDLINE 编制索引的关键词。

• MeSH Major Topic 包含了所有在 MeSH 中被标记为非常重要的术语。

• Publication Date 用于搜索文章发表、序列公布或提交的日期。

• Modification Date 资料被收入 Entrez 的日期。

• Page Number 发表文章的页码。

• Property 告诉用户引文包含的序列类型。

• PubMed ID 给定引文的 PubMed 标志符。

• Organism 用于搜索与蛋白或核苷酸序列条目有关的生物体的名称（包括普通名称和学术名称）

• Protein Name 用于搜索与一个序列数据相关的蛋白质的名称。

• SeqId 给定序列的串标志符。

• Substance 搜索在 Chemical Abstract Service（CAS）上登记的化学物质的名称。

• Title Words 用于搜索仅出现于某一记录的标题里的词。

• Text Words 用于搜索与某一给定记录相关的说明（“free text”)。对于蛋白和核苷酸序列，则包括给定序列的定义、说明、命名和描述。在 MEDLINE 条目中，包括给定记录的标题和摘要。

• Volume 搜索所要文章的卷号。

如果在某一特定的搜索栏内找不到所要的资料，那么用 All Fields 或者 Text Words 重复搜索将会有所帮助。在 Entrez 中交叉符用 AND，仅查找被 AND 分开的包含所有给定术语的相关信息资料。Entrez 将连结符指定为 OR，可使用户查找到包含任何一个给定术语的相关文献。最后，差别选项是 BUTNOT，使用户能查找到包含了上一个术语而不包含下一个术语的所有文献。搜索成功后，用户可以在文献清单中进一步检索，直至达到要求为止。搜索结果的清单按文件从最近到过去的时间顺序出现，用户既可检索所有文献也可从所查清单中选择最相关的报告。以下是几种针对不同的检索文档的不同的查看格式：

PubMed 文章可以用 Citation、Abstract、MEDLINE 或者 ASN. 1 典型格式查看。

• Citation 格式能显示文章的标题、摘要、MeSH Terms 和主要信息。
• Abstract 格式只显示文章的标题和摘要。
• ASN. 1 是应用于 PubMed 文章的一种特殊格式。
• MEDLINE 则以 MEDLARS 格式显示文章。

GenBank/GenPept、Report、ASN. 1、Graphic View 及 FASTA 是一些用于查看蛋白或核苷酸序列记录的格式。

• GenBank/GenPept 是标准的 GenBank 或 GenPept 数据库文件。
• Report 允许用户以 GenBank Report 格式查看序列记录。
• Graphic View 使用户能查看序列条目的图表资料包括排列信息。
• FASTA 格式对于给定条目的进一步分析最为有用。

许多排列工具要求用户在 FASTA 格式下输入所关心的序列顺序。Structure Summary 和 ASN. 1 格式用于查看结构信息，Structure Summary 格式用于获取给定分子结构资料的概要。例

如，晶体蛋白结构，这种查看格式可使用户获取关于给定结构的解析度、作者资料、提交日期、复杂化的配基以及其他基本信息。这种格式还允许用户查看分子的三维形式。Graphic View 也用于查看基因组记录资料。

所有被提及的格式都可作为文件保存在用户文档中。三种主要保存格式是：Text、HTML 和 MIME 。如用户拥有 GenBank 的 MIME 浏览器，MIME 格式是特别有用的。否则，输出文件必须保存在 Text 或 HTML 格式中应用。如用网页浏览器查看，则可用 HTML 格式。Text 格式缺乏 HTML 标签，但可用标准文字处理软件，如使用 Microsoft Word 来查看。

4. BankIt

BankIt 是 GenBank 通过互联网进行操作的序列提交服务器。它允许用户通过界面友好的网页浏览器将新的序列提交到 GenBank。该序列及所有相关信息被传递到提交信箱并送到 GenBank。GenBank 工作人员再与提交当事人取得联系，给该序列指定序列号。

5. OMIM（Online Mendelian Inheritance in Man）

这是关于人类基因和基因疾病的数据库，由 Victor A. Mckusick 博士和他在 Johns Hopkins 大学的同事以及一些其他捐助人共同维护。The OMIM Morbid Map 也由该站点支持，在遗传疾病的基础上绘制基因位点图谱。Entrez、GDB、The Davis Human/Mouse Homology Map、the Online Mendelian Inheritance in Animals（OMIA ）、the Human Gene Mutation Database (HGMD)、the Alliance of Genetic Support Groups、the Cedars-Sinai Medical Center Genetics Image Archive、the Jackson Laboratory、RetNet (retinal genetic disorders)、HUM-MOLGEN 以及 the locus-specific mutation databases 都是一些可在 OMIM 得到的资源。该站点特别适应那些关注遗传疾病的医生和医学科研人员。若要给 OMIM 的图象和文章以最适当的诠释，那么对科

学概念及其研究进展的坚实理解是十分必要的。

6. Taxonomy

NCBI 的 Taxonomy 主页是包含一些生物的普通名称和科学名称的生物分类数据库，这些生物或多或少都包含有一些序列信息。该服务器允许用户得到种属的遗传信息，观察相关和不太相关的种属间是如何联系的。进化树是这种联系的代表。这些关联是基于相似的蛋白或核苷酸序列上的。该主页还与 NCBI 其他服务器链接（如 Structure 和 PubMed）。

7. Structure

NCBI 的 Structure 主页支持与结构分析相关的分子模型数据库（MMDB）和各种各样的软件工具。MMDB 的信息是从 the Brookhaven Protein Data Bank（PDB）中获得的，包括重要生物大分子 X 线晶体衍射或核磁共振（NMR)的结果。Cn3-D是 NCBI 下为 MMDB 服务的结构可视软件，可在 Entrez/ Cn3-D FTP 站点获得。Structure 也提供如 PKB 和 Threading 的搜索工具，该软件可通过 FTP 站点获得，并要求起用 Splus。该站点的 Entrez/PubMed 链接便于查找所关注的分子的应用及相关信息的搜索工作。

8. Books

这是 NCBI 与作者和出版机构合作建立的书籍目录查询服务项目，称为"Bookshelf"，主要收集生物医学书籍。在 PubMed 的条目下，有相关书目的链接。经过 Entrez 的查询可得到每个 PubMed 条目中有标注"Books"的链接，点击该链接可以查询到与每个 PubMed 条目在主题上相关的生物医学书籍目录。另外，还可以通过 Entrez 直接输入查询的主题词，检索相关的书目。

NCBI 的 Hot Spots：

Hot Spots 是 NCBI 主页提供的最常用的一些数据库和搜索工具的链接，包括以下内容：

（1）Cancer Genome Anatomy Project

(2) Clusters of Orthologous Groups
(3) Coffee Break
(4) Electronic PCR
(5) Gene Expression Omnibus
(6) Genes and disease
(7) Human genome resources
(8) Human map viewer
(9) Human/mouse homology maps
(10) LocusLink
(11) Malaria genetics & genomic
(12) ORF finder
(13) Reference sequence project
(14) Retrovirus resources
(15) Serial analysis of gene expression
(16) SKY/CGH database
(17) Trace archive
(18) UniGene
(19) VecScreen

1. Cancer Genome Anatomy Project

肿瘤基因组解剖计划（Cancer Genome Anatomy Project，CGAP）由美国国立癌症研究所（National Cancer Institute，NCI)建立和管理，与 NCBI 有密切的合作。肿瘤基因组解剖计划旨在得到用于解码肿瘤细胞分子解剖的信息和工具。它由几个互补的数据库组成： Human Tumor Gene Index、Molecular Fingerprinting、Cancer Chromosome Aberration Project、Genetic Annotation Initiative 及 Mouse Tumor Gene Index 等。其目的是研究正常细胞、癌前病变细胞和肿瘤细胞的基因表达谱，以便最终能够促进肿瘤病人的检测、诊断和治疗。 CGAP 网站可提供人类、小鼠正常和肿瘤组织的基因组数据，包括表达序列标签

（ESTs）、基因表达谱、单核苷酸多态性（SNPs）和细胞遗传学信息等等。它还提供了查询和分析数据的信息学工具以及由该计划开发出来的试剂的使用方法和资源信息（http://cgap.nci.nih.gov/）

NCBI 与 NCI 的合作工作呈流水线方式，首先产生大量的表达序列标签(EST)，然后储存入 dbEST 数据库，经过计算机软件对 EST 的分析和总结，将其整和成 UniGene 和 HomoloGene 数据库，然后通过加入手工注解并形成 LocusLink 数据库。NCBI 站点每周贴出文库报告，对 NCI 构建的 EST 文库提供文库分布的信息，CGAP 利用这些信息来指导将来的文库构建和测序工作。NCBI 创建了第一个公共 CGAP 站点，并为 CGAP 设计所有的分析工具。NCBI 开发了在线工具——数字差异显示（Digital Differential Display），用来在 cDNA 文库之间比较计算机产生的基因表达谱。

基因表达系列分析（Serial Analysis of Gene Expression，SAGE)是 CGAP 中分析基因表达的重要方法，这种方法的原理将在本书有关章节中介绍。 NCBI 建立了序列数据库用来储存 SAGE 产生的表达谱数据。SAGE 数据也收入 Gene Expression Omnibus 数据库中。

CGAP 数据库中提供的资源包括 8 个方面，下面分别将其作一简单的介绍：

（1）Genes

CGAP 对于每个基因都开发了一个 Gene info 页面，每个页面提供了该基因与 NCBI 和 NCI 数据库相关条目的链接信息，这些信息包括与 UniGene、LocusLink、OMIM、DTP search、cDNA Libraries、Cluster Assemblies 和 SNPs 等数据库的链接；细胞遗传学定位和 Mitelman 断裂点信息；蛋白相似性、人类和小鼠的同源组、IMAGE（Integrated Molecular Analysis of Genomes and their Expression）协议来源的序列链接、全长 MGC（Mammalian

Gene Collection ）克隆链接和 Gene Ontology 功能分类。CGAP Genes 工具列表：

- Gene finder：按照特定标准查找单个或多个基因的工具。
- Gene Ontology Browser ：通过分子功能、生物学过程和细胞组分对人和小鼠的基因分类。
- Nucleotide BLAST 通过 CGAP 界面，查找给定核苷酸序列的代表基因。
- Lists of Candidate，Validated，and Confirmed SNPs 包含单核苷酸多态性的基因信息。
- CGAP SNP Index：通过基因名称、符号和 GenBank 序列号查找代表 SNPs。
- SNP Gene Viewer 将人类 SNPs 定位于参考序列和 MGC 序列，预测蛋白编码的变化。

（2）Chromosomes

瑞典的 Mitelman、Mertens 和 Johansson 博士曾系统地总结了经常发生的肿瘤相关染色体畸变，并建立了 Mitelman 肿瘤染色体畸变数据库（Mitelman Database of Chromosome Aberrations in Cancer ）（http://cgap.nci.nih.gov/Chromosomes/Mitelman），这个工作发表在 1997 年 4 月出版的 Nature Genetics 杂志特刊上，题目为“A genome-wide map of chromosomal breakpoints in human cancer”。目前，所有常见的肿瘤染色体畸变可以在 NCBI CGAP 的站点 Recurrent Chromosome Aberrations in Cancer（http://cgap.nci.nih.gov/Chromosomes/RecurrentAberrations ）进行交互式查询。

NCBI 还与 CGAP 的一个分支计划——肿瘤染色体畸变计划（Cancer Chromosome Aberration Project，CCAP ）紧密合作。CGAP 目前正在制作跨度为 1—2Mb 的人染色体 BAC 克隆，这些克隆都是通过荧光原位杂交（Fluorescent In Situ Hybridization，FISH）的方法定位的，这些已经定位的 BAC 克隆可以提供给其他

研究组织。

另外，CGAP 还提供了 Genetic and Physical SNP Maps，该图谱可以显示每条染色体上预测的和经证实的 SNPs 的遗传和物理位点。

（3） Tissues

CGAP Tissues 资源提供来源于组织的基因表达信息。Library Finder 工具可以帮助查找组织特异性的文库。该资源包括了几种分析组织基因表达分析工具：

- Gene Library Summarizer（GLS）在特定的 cDNA 文库中查找所有基因。
- cDNA xProfiler ：用于在两个 文库之间比较基因表达。
- Differential Gene Expression Displayer （DGED）：用于在两个 cDNA 文库之间比较基因表达的统计学差异。
- SAGEmap xProfiler：xProfiler 程序比较一个 cDNA 在不同 cDNA 文库中的表达情况。
- SAGEmap Virtual Northern ：以图形形式显示一个基因在不同 cDNA 文库中出现的频率，代表其表达丰度。

（4） Pathways

Pathways 资源包括了 公司和 代谢途径和信号转导途径图谱的链接，网站提供了代谢途径和信号转导途径的精美图片和图例，见图 3-3。

（5） Tools

本窗口是 CGAP 基因表达分析工具的汇总，方便用户从网上得到这些工具。

（6） Methods

该链接提供了 CGAP 在分析基因表达中使用的组织准备、cDNA 文库构建的详细方法。另外还包括了 CGAP 应用的激光捕获显微切割的方法介绍。

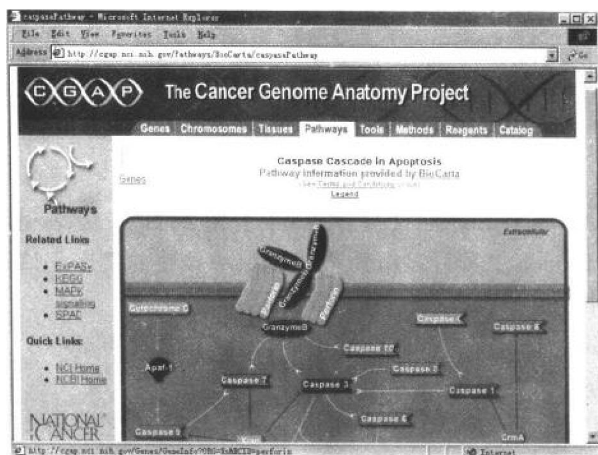


图 3-3　CGAP 提供的 BioCarta 中 Caspase 分子相关凋亡信号转导通路图（http://cgap.nci.nih.gov/Pathways/BioCarta/caspasePathway）

（7）Reagents

该站点包括 CGAP 对研究单位提供的 cDNA 克隆、BAC 克隆和 cDNA 文库的目录以及查询方法。

（8）Catalog

Catalog 列出了 CGAP 提供的资源目录，包括人和小鼠的 Chromosomes、Clones、Genes、Tissue and Librarys 和 SNPs 等方面的资源。

CGAP 也通过光谱核型（Spectral Karyotyping，SKY）和比较基因组杂交（Comparative Genomic Hybridization，CGH）的方法来确认肿瘤中的染色体畸变，并且专门有 SKY/CGH 数据库（http://www.ncbi.nlm.nih.gov/sky/）来收录这些信息。

2. Clusters of Orthologous Groups

该链接内容在第四章第四节有详细介绍。

3. Coffee Break

Coffee Break 收集了一些短小精悍的综合报告，它们的内容涉及到利用 NCBI 的工具发现的最新生物医学进展。每篇文章大约是 400 字左右，且附有精美的相关图片（图 3-4），及在研究工作中如何应用 NCBI 工具和资源的超级链接。其行文活泼，但不失严谨和实用性。

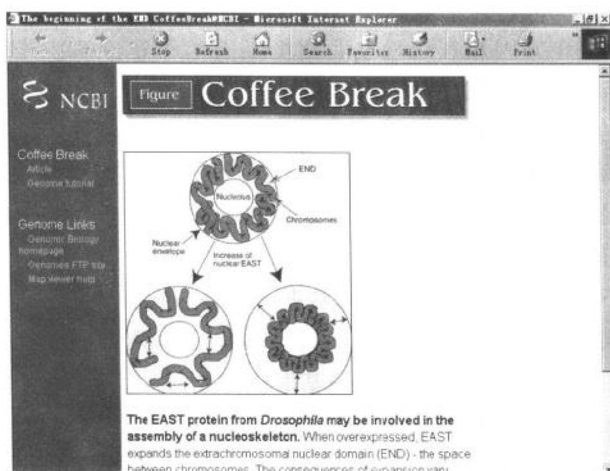


图 3-4 Coffee Break 的界面

4. Electronic PCR

电子 PCR（electronic PCR）是一种可查找 DNA 序列中是否含有序列标签位点（STS）的软件。在基因组计划中非常重要的一项工作，就是确定应用于基因组作图的标志——序列标签位点，测序工作完成后就可以利用电子 PCR 查询其中是否包含 STS 序列，这样就将基因组测序和作图的工作联系起来。另外，电子 PCR 还可以检测根据 STS 序列设计的 PCR 引物是否完全与 STS 序列匹配、方向是否正确以及 PCR 产物的分子量。

5. Gene Expression Omnibus

Gene Expression Omnibus 收集基因表达的有关数据，包括

来自于不同平台的数据，如：利用微阵列（microarray)基因表达数据、高密度寡核苷酸阵列（high-density oligonucleotide array，HDA）、杂交滤膜（hybridization filter）和基因表达系列分析(Serial Analysis of Gene Expression，SAGE ）等方法得到的基因表达数据。

6. Genes and disease

NCBI 基因和疾病站点收集由基因变异导致的遗传性疾病，数据库中大部分数据是已经研究清楚的由单个基因突变直接引起的遗传性疾病。目前共收集了 73 种遗传性疾病，分为六大类，每种疾病的遗传位点都在染色体上以图形形式标出，并与 PubMed、LocusLink 和 OMIM 数据库有链接(http://www.ncbi.nlm.nih.gov/disease/)。

下面以肿瘤为例，介绍本站点可获得的数据资源。 NCBI 的基因与疾病站点(http://www.ncbi.nlm.nih.gov/disease/Cancer.html）中列出了多种肿瘤相关基因的变异：

- 乳腺癌：BRCA-1（17 号染色体）、BRCA-2（13 号染色体）
- Burkitt 淋巴瘤：myc（8 号染色体）
- 慢性髓系白血病：BCR（22 号染色体）、ABL（9 号染色体）
- 结肠癌：MLH1（3 号染色体)、MSH2 和 MSH6（2号染色体）
- 肺癌：SCLC1（3 号染色体）
- 恶性黑色素瘤：CDKN2（9 号染色体）
- 多发性内分泌肿瘤：MEN1（11 号染色体）
- 神经纤维瘤：NF2（22 号染色体）
- p53 肿瘤抑制基因：位于 17 号染色体
- 胰腺癌：DPC4（SMAD4），位于 18 号染色体
- 前列腺癌：HPC1（位于 1 号染色体）

• ras 癌基因：HRAS（位于 11 号染色体）

• 视网膜母细胞瘤：RB1（位于 13 号染色体）

• von Hippel-Lindau 综合症：VHL（位于 3 号染色体）

网站中各种肿瘤相关基因都有与染色体定位、GenBank 序列、PubMed 文献和 OMIM 数据库的链接。

7. Human genome resources

人类基因组资源主页提供了人类基因组计划相关内容的链接，包括了人类基因组测序、作图、遗传变异和基因表达的核心研究资源，可以以图形的方式利用 Human map viewer 直接搜索人染色体的作图、测序数据。其图形界面非常直观，便于科学工作者查询。

8. Human map viewer

Human map viewer 是查询人类染色体相关数据的图形界面的程序，它以不同的染色体图形的方式来显示（图 3-5），点击相关位点可以显示该位点的作图和测序数据。

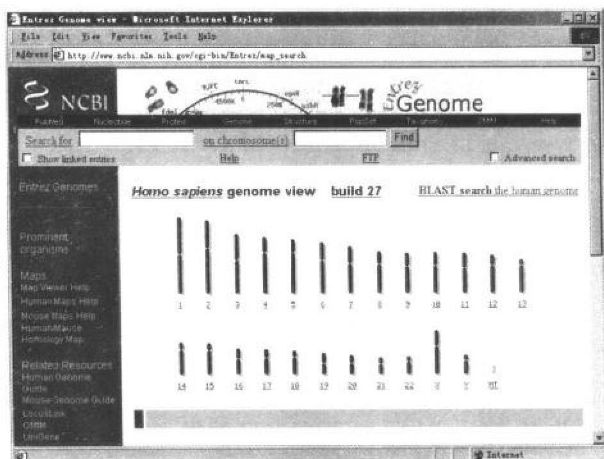


图 3-5 Human map viewer 显示的人类染色体资源，点击每条染色体可以得到详细的信息。

9. Human/mouse homology maps

此站点列出了人和小鼠同源 DNA 片段的遗传位点，共收录了 6453 个位点。并与 GeneMap'99、OMIM 和 Jackson Laboratory 的 the Mouse Genome Database 有链接。

10. LocusLink

LocusLink 提供了一个简单的界面，可用来查询人类基因或遗传位点的准确信息，包括基因的官方术语、别名、序列号、表型、EC number、MIM number、UniGene 簇、作图信息等。它与一些相关的 Web 站点链接以提供完善的信息，包括：Human Gene Nomenclature Committee (HGNC) (http://www.gene.ucl.ac.uk/nomenclature/)、the Genome Database (GDB; http://gdbwww.gdb.org/)、the Human Gene Mutation Database (HGMD; http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html)、GeneCard (http://bioinfo.weizmann.ac.il/cards/) 和 GeneClinics (http://www.genecli nics.org/)。

11. Malaria genetics & genomic

该站点提供了疟原虫 (Plasmodium falciparum) 的遗传学和基因组学的信息，包括基因组图谱、连锁标记和遗传学研究结果 (http://www.ncbi.nlm.nih.gov/Malaria/)。

12. ORF finder

ORF Finder 是查找 DNA 序列中开放读框的软件，详细内容在第四章第二节有介绍。

13. Reference sequence project

RefSeq 与 GenBank 不同，它是另一类型的基因数据库。它只收录有全编码区的或功能已有一定研究的基因。RefSeq 的记录分为两种：临时 (provisional) 记录和已编辑 (reviewed) 记录。它首先发布临时记录，包括来源于 GenBank 的各种注解，并增添了基因和蛋白的名称、PubMed 链接、摘要文本和来自 LocusLink 的基因作图和染色体数据。经过专家审定和修改后生成已编辑记录，

包括了更多的信息：a. 经过整理后延长的基因 5'和 3'非翻译区序列。b. 更多的 mRNA 和蛋白特征。c. 发表的相关文章。d. 描述基因特征的一段摘要。RefSeq 记录与 OMIM、PubMed、GenBank 和 UniGene 都有链接，RefSeq 记录用'NM_******'和"NP_******"表示。RefSeq 可通过基因或蛋白名称、序列号以及序列同源性来查询，Entrez 和 LocusLink 都支持用文本来查询 RefSeq，BLAST 也可用 RefSeq 的序号来作同源比较，但应在序号前加前缀"ref"（http://www.ncbi.nlm.nih.gov/LocusLink/refseq.html）

14. Retrovirus resources

反转录病毒基因型分析工具（Retroviral genotyping tools），此工具的目的在于揭示反转录病毒遗传多样性的特点，追溯其流行病学规律。NCBI 的基因型分析工具主要是利用 blastn 程序对用户提供的反转录病毒序列和一系列相关序列进行相似性比较并划分亚型，相当于一个多序列对齐的过程。对于 HIV 来说，这个工具就是为开发疫苗而设计的。HIV 专用的亚型分析工具中包括了 HIV 的 M 组 A-J 亚型和 O 组、N 组的基因组参考序列。另外，还包括了 SIV、HTLV、STLV 几种病毒的专用分析工具。还可以利用 NCBI 提供的 Cn3D 工具对照 MMDB 观察病毒分子三维结构的变化，并与多种病毒基因组的图谱链接，与 National Institute of Allergy and Infectious Disease（http://www.niaid.nih.gov/research/daids.htm）、HIV Sequence Database（http://hiv-web.lanl.gov/）、Sanbi HIV Africa（http://ziggy.sanbi.ac.za/hivafrica/）、Stanford HIV RT and Protease Gene Database（http://hivdb.stanford.edu/hiv/）等相关数据库链接。

15. Serial analysis of gene expression

基因表达系列分析（Serial Analysis of Gene Expression，SAGE）是对某一组 mRNA 中基因表达定量检测的一种方法。如果在一群 mRNA 中的每一个 mRNA 的同一位置取 9-10bp 的一

段序列，（从统计学上说这样的序列可以特异地代表着 95%的人类基因),每一种 9-10bp 的序列的拷贝数则代表基因表达的拷贝数，也就可以说明基因表达活性的高低。 Johns Hopkins 大学研究小组利用了这种方法，首先提取组织 RNA ,得到 cDNA ,同时用生物素标记 cDNA 末端。随后，用一种限制性内切酶切割 cDNA，这样就可以分离到酶切位点 3' 端的序列。用另一种酶再切割 cDNA 片段，去除带生物素的 3 ' 端后用 PCR 扩增每一个标签片段 ,将 30-50个不同的标签片段连成一个单一的 DNA 分子。最后克隆并测序这些分子，这样酶切位点 3 ' 端的序列出现的拷贝数就代表了基因表达活性的高低。基于以上原理， NCBI 建立了 SAGE 数据库，此数据库主要依靠 UniGene 簇来建立，可以用 SAGE 标签来查询 UniGene 簇，也可以用 UniGene 簇来查询 SAGE 文库中代表这一 UniGene 簇的 SAGE 标签出现几率（http://www.ncbi.nlm.nih.gov/SAGE/）

16. SKY/CGH database

该数据库收集光谱核型（Spectral Karyotyping，SKY ）和比较基因组杂交（Comparative Genomic Hybidization，CGH）的数据。光谱核型是用不同颜色的荧光来标记染色体，以显示人和小鼠的所有染色体，使得染色体畸变更容易观察到。比较基因组杂交是利用肿瘤或参照 DNA 的探针与正常或肿瘤的染色体杂交，用来发现在肿瘤基因组中 DNA 拷贝数的改变（http://www.ncbi.nlm.nih.gov/sky/）

17. Trace archive

Trace archive 用来储存多种生物基因组计划中所有序列的原始资料，研究者可以根据自己的兴趣按照提供的查询方法得到最原始的基因组测序结果。

18. UniGene

EST（expressed seqeunce tag，EST）称表达序列标签，是从 cDNA 克隆中随机挑选出来进行一次性测序的结果。一般长约

200-500bp，通常作为基因的标志。至 2002 年 1 月 GenBank 中收集的 EST 有 10,069,598 个序列。由于 cDNA 文库的复杂性和测序的随机性，有时多个 EST 代表着同一基因或基因组。通过对 EST 的分析将其归类而形成 EST 簇（EST cluster），每一个 EST 簇代表着一个特定的基因，即 UniGene。而 UniGene 数据库收集了大量的 EST 簇，并与相关信息链接，如：表达的组织类型、染色体作图、表达的蛋白等。UniGene 簇用 Hs. ******来表示。目前在人类 UniGene 数据库（http://www.ncbi.nlm.nih.gov/UniGene/）中，用超过 150 万的 EST 构建了 83000 个 EST 簇，代表了大部分的人类基因。EST 簇的 3’非编码区的序列还可以转换成序列标签位点（sequencetagged sites，STS）的序列帮助基因组作图工作。如果采用 DNA 芯片技术研究大量基因的表达情况，还可利用大量的 EST 簇的序列来设计芯片。UniGene 可通过基因名称、染色体、cDNA 文库、序列号和普通文本等格式来查询，还可通过 FTP 下载。

19. VecScreen

VecScreen 是查找序列中是否包含有载体序列的软件，可以发现序列中是否有载体序列污染。

二、欧洲生物信息学研究所（European Bioinformatics Institute, EBI）

EBI 是 European Molecular Biology Laboratory（EMBL）的分部，位于英格兰的 Hinxton。由 14 个欧洲国家和以色列支持 EMBL 和其分部。EBI 的主要目标是从事研究并为全球科学界提供生物信息学资源。1994 年 9 月，EBI 承担了所有此前由德国海德堡（Heidelberg）的 EMBL’s Data Library 掌管的项目。EBI 可以比作美国的 NCBI，是欧洲主要的生物信息服务机构（图 3-6）。其任务和目标与 NCBI 相同，包括：

- 生物信息学技术跟踪
- 研究开发生物信息软件

- 对订阅用户提供培训和支持
- 相关的生物信息服务

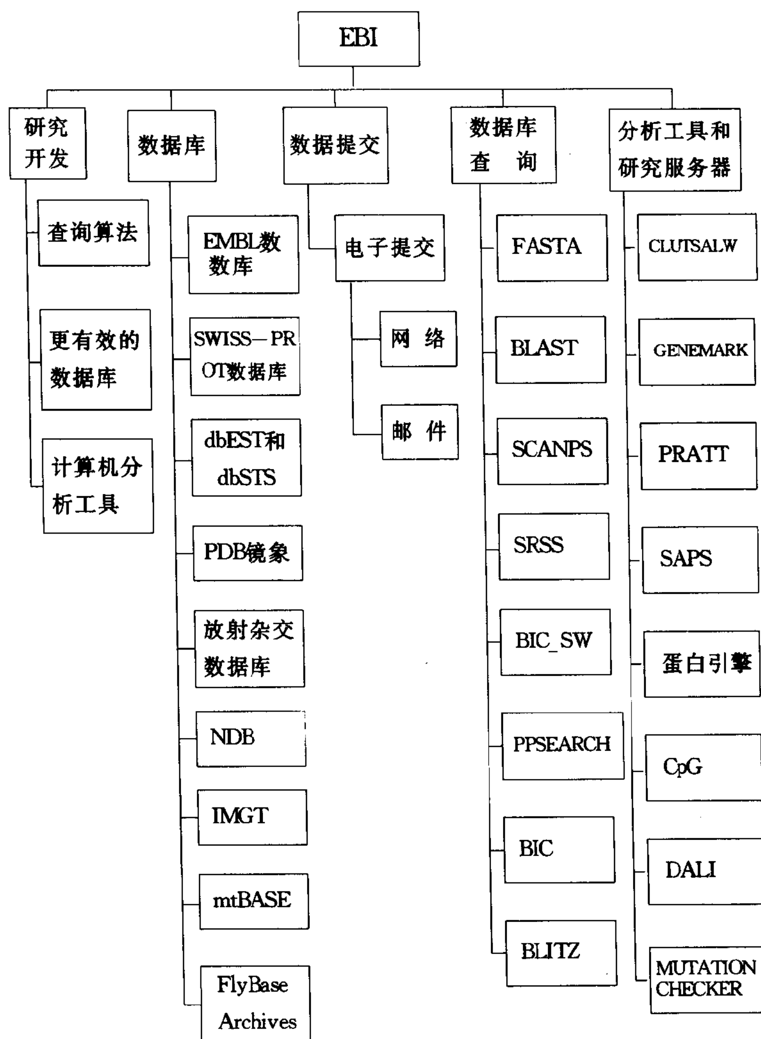


图 3-6 欧洲生物信息学研究所的结构图

EBI 员工包括计算机专家、分子生物学家、数学家、生化学家、医学研究人员和结构生物学家。雇员们通力合作，运用数学和计算机工具研究疾病的分子基础。 EBI 的科研工作涉及生物信息领域的许多方面。主要的科研任务包括：

- 开发更为强大的比较运算算法
- 创立更为详尽，界面友好的网际信息系统
- 设计更高效的数据库

EBI 提供的服务：

- 数据库
- 数据提交
- 查询数据库及相似性搜索（如 FASTA 和 BLITZ）
- 在线应用程序
- FTP 存档文件
- 研究和开发

EBI 的数据库：

由EBI 网页服务器支持的主要数据库和分析工具见表 3-1。

表 3-1 EBI 支持的数据库和分析工具列表

数据库名称	说明
核苷酸序列数据库	
Simple Queries	应用 SRS（Simple query retrieval）检索系统简单检索数据库
EMBL database	EMBL 核酸数据库
EMBL-Align database	EMBL-Align 多序列排列数据库
Ensembl	自动注解的真核生物基因组
DbEST and dbSTS Queries	查询 dbEST 和 dbSTS 的工具
EMEST	EMBL 的 EST 序列数据库
EuroGeneIndexes	EST 经排列后总结成基因簇的数据库
MitBase	线粒体 DNA 数据库
IMGT	ImMunoGeneTics 数据库
EDGP	欧洲果蝇基因组计划数据库

续表

数据库名称	说明
Parasites	寄生虫基因组数据库
Mutations	序列变异数据库计划
Genomes Server	由 EBI 完成的基因组数据总缆
Genome MOT	基因组监测表
蛋白质序列数据库	
SWISS-PROT、TrEMBL、InterPro、etc	SWISS-PROT、TrEMBL 和 InterPro 等蛋白序列数据库
CluSTr	将 SWISS-PROT 和 TrEMBL 中的蛋白自动分类为相关的分组
序列结构分类数据库	
DSSP	二级结构数据库
HSSP	经同源分析方法得出二级结构数据库
FSSP	基于结构比较的折叠分类数据库
DALI	蛋白质结构、结构域词典
3Dee	蛋白质结构域定义数据库
大分子结构数据库	
EBI-MSD	EBI 大分子结构数据库，包括 PDB 搜索工具
NDB：EBI Mirror	EBI 的镜像站点，暂时未提供
序列作图数据库	
RHdb	放射杂交数据库
GenomeMaps 98	人类基因组图谱 98
档案（Archives）	
Software Biocatalog	分子生物学软件的路径
FlyBase Archives	果蝇基因组档案库
EBI ftp server	EBI 的数据库和软件 FTP 下载服务器
BioWurld	互联网中的生物信息学和分子生物学资源

下面选择其中比较重要的六个数据库讨论，其他 EBI 的服务和工具可以通过其网址 http://www.ebi.ac.uk/获得。

1．EMBL 核苷酸序列数据库

EMBL 是一个内容广泛的核苷酸（DNA 和 RNA）序列的数据库。其核苷酸序列来源于很多渠道，有些源自科学文献和专利申请，但大部分则是由研究者或测序小组直接提交的序列原始资料。该数据库与美国 NCBI 的 GenBank 核苷酸序列数据库和日本的 DNA 数据库（DDBJ）合作，通过软件程序每天交换数据。EMBL 数据库与这两个数据库保持联系，可不断地更新内容，从而使 EMBL 能为全球科技界提供所有公共范围内已知的核苷酸序列资料。另外，EMBL 与众多的基因组测序小组合作，可以大规模地获得核苷酸序列。

（1）EMBL 核苷酸序列文档的信息类型：

· 序列

· 序列的简单描述

· 序列来源（序列所属物种）

· 文献目录及引文信息

· 序列中编码区的位置（如信号序列、 α 链及 β 链等）

· 序列中有生物学意义的位点（EST 是单向序列，主要来源于随机克隆，较少有已知的功能和生物学信息。由测序小组提交的序列是经细致注释的，含有报告这些条目的研究者的深入研究结果）。

（2）EBI 提供的已完成基因组资源（截止到 2002 年 1 月，http://www.ebi.ac.uk/genomes/）

· 原生质：11 种

· 细菌：63 种

· 真核生物：5 种

· 细胞器：201 种

· 噬菌体：90 种

· 质粒：246 种

· 类病毒：37 种

· 病毒：628 种

2. SWISS-PROT 蛋白序列数据库

该数据库由日内瓦大学和 EBI 的 EMBL 数据库联合维护。EMBL 核酸序列数据库的编码 DNA 序列翻译成氨基酸序列后，文档保存在 TrEMBL 数据库中。由肽段测序计划中得出的蛋白序列直接提交到 SWISS-PROT，并指定序列号。TrEMBL 数据库包含两个部分：SP-TrEMBL 和 REM-TrEMBL，SP-TrEMBL (SWISS-PROT TrEMBL）数据库中的条目最终应当整和到 SWISS-PROT 中，且每个序列都指定了一个 SWISS- PROT 序列号。而一些 EBI 暂不想放入 SWISS-PROT 的条目收录在 REM-TrEMBL 数据库中，没有指定 SWISS-PROT 序列号。SWISS-PROT 为非冗余数据库，它在 2001 年 10 月公布的 SWISS-PROT Release 40.0 版本包含了 101,602个序列条目。而 2001 年 12 月公布的 TrEMBL Release 19 版本包含了 636,825 个条目。如它与 EMBL 数据库间的交互参考可使用户获得核苷酸序列信息。它同时拥有来自 PDB 及 PROSITE 数据库的参考资料，PDB 参考资料只能在已知三维结构的序列条目中找到，PROSITE 的参考资料也只能在包含 PROSITE Motif 的序列条目中找到。

3. 放射杂交数据库（Radiation Hybrid Database，RHdb）

RHdb 是收录用于构建放射杂交图谱的原始数据的数据库，包括 STS 数据、分值、实验条件和多方面的交叉参考数据。2001 年 1 月公布的 RHdb Release19.0 版本包含约 229 个实验条件、92 个图谱的三个物种（包括人、小鼠和大鼠）的 133,239 个放射杂交条目（其中有 106,574 个 STSs）。

放射杂交图谱是根据放射杂交矢量分值（RH score vector）计算后构建的染色体图谱，是另一种遗传图谱。由于放射杂交图谱可以包括非多态性标记，对于遗传图谱的完善是不可缺少的补充，且可以对未澄清的多态性 STS 簇排序。简单地说，两个标记的矢量分值越相似，它们在染色体上的位置就越接近。国际上的合作

研究计划产生了大量的人、小鼠和大鼠的杂交数据，这样就可以构建准确的 STS 图谱。它对于研究人类多因素遗传疾病有重要价值。

4. dbEST 和 dbSTS

这是 NCBI 的 EST 和 STS 数据库的镜像数据库。该数据库条目主要包括：表达序列标签(expressed sequence tags，ESTs）——单向测序的 cDNA 序列条目，序列标签位点（sequence tagged sites，STSs）以及短基因组标记序列（short genomic landmark sequences），均由 NCBI 加以维护。 EBI SRS 界面可用于搜索 dbEST 和 dbSTS 数据库。

5. PDB（Brookhaven 镜像站点） Protein Data Bank

PDB 数据库收集所有已知的三维结构信息，最初由 Brookhaven 国家图书馆负责维护，从 1999 年 7月起，由 the Research Collaboratory for Structural Bioinformatics（RCSB）接管，其网址也相应地该为 http://www.rcsb.org/pdb/ 。由美国国家科学基金会（U. S. National Science Foundation)、国家公共医疗科学协会（National Institute of General Medical Sciences） 、国家医学图书馆（National Library of Medicine ）和美国能源部(U. S. Department of Energy）共同提供资金支持。

（1）PDB 维护的结构类型：

- 蛋白质
- 蛋白质＋核苷酸序列（如 DNA）
- 蛋白质一金属复合物
- 蛋白质一抑制剂复合物

（2）三维结构的确定方法包括：

- 核磁共振（NMR）
- X 线晶体衍射技术

（3）两种技术的差异：

- 磁共振是在溶解状态下检测分子的结构。这项技术能得出

大量动力学资料，能观察到水溶液（溶解）状态下的分子行为。使我们能通过分子的结构特性了解其功能特性。

- X 线晶体衍射技术研究分子的三维静态图象。分子结构决定晶体形态，因此，通过该技术测定的分子结构缺乏动力学资料。换句话说，我们无法知道在溶解状态、自然状态下分子的行为。

既然溶解的分子结构经核磁共振可以得到更多的信息，为何不应用这种方法研究所有的分子结构？这是由于经核磁共振检测溶解的分子结构会受到分子大小的限制。许多蛋白都超出了能够检测的范围。因此，必须利用可观测较大分子结构的其他技术（如 X 线晶体衍射技术）。

（4）PDB 文档中信息类型：

- 运用核磁共振或 X 线晶体照相术确定的原子间关系
- 引用的文献
- 一级结构信息（如氨基酸序列）
- 二级结构信息（如 α 螺旋、β 片层结构）
- 晶体学结构因素和核磁共振实验数据

6. IMGT 数据库（The International ImMunoGeneTics Database）

IMGT 数据库创建于 1989 年。当时这项开发工作由法国 Montpellier II 大学完成。这是一个核苷酸数据库，其中包括许多属于免疫球蛋白超家族的重要的免疫学相关基因。免疫球蛋白超家族中的大部分分子涉及免疫识别和免疫应答。 T 细胞受体（TCRs）、免疫球蛋白（Ig）和主要组织相容性复合物（MHC）分子都是典型的免疫球蛋白超家族成员。由于该数据库具有高水平，并且信息分布简单，它对医学研究有很大帮助。其中包括对自身免疫性疾病、AIDS 、白血病、淋巴瘤和骨髓瘤等疾病的研究，并且对治疗方法、抗体工程相关的生物技术、免疫应答中的基因多样性和基因进化等方面研究均有重要的提示帮助。图 3-7 是

IMGT 数据库的结构图（http://imgt.cines.fr：8104/）。

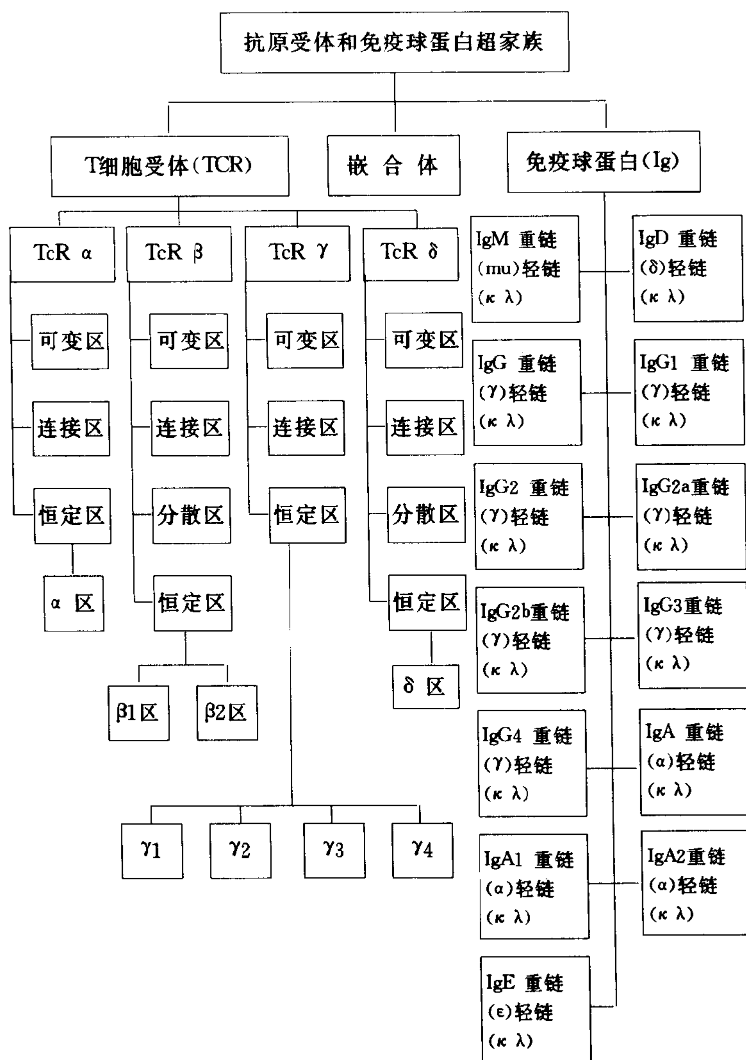


图 3-7 IMGT 的分类示意图

（1）可从 IMGT 检索的信息类型：

- 核苷酸序列
- 蛋白序列
- 序列排列方式
- 等位基因、多态性和 STS 信息
- 基因图谱和遗传学数据
- 结构数据
- 寡核苷酸引物
- 与疾病的关系

（2）储存于 IMGT 数据库的分子大都是有免疫学意义的。根据分子特性，IMGT 数据库数据分别贮存于两个不同的数据库。

- LIGM-DB 是贮存免疫球蛋白及 T 细胞受体分子的数据库。LIGM-DB 表示 the Laboratoire d'ImmunoGenetique Moleculaire。至 2001 年 12 月已经拥有人和其他脊椎动物的 53936 个免疫球蛋白和 T 细胞受体的核酸序列条目，序列具有全面的注释。
- MHC/HLA-DB 这是关于主要组织相容性复合物分子的数据库。在人类，这类分子指人类白细胞抗原（HLA)。至 2001 年 10 月该数据库包含 1468 个人类 MHC 等位基因序列。

（3）服务对象和可能的用途：

- 医学研究人员（如 HIV/AIDS 研究、癌症研究和自身免疫性疾病的研究人员）
- 治疗方法和免疫组织化学研究（如抗体的制备、某些免疫疗法、移植免疫中的对抗剂等）
- 进化生物学家和生物信息学家、进化学以及与其相关的基因组多样性研究。在不同种属中分子间的关系，可能是寻找致病相关基因的强有力工具。

（4）IMGT 数据库的主要合作者：

• LIGM：法国蒙彼利埃的蒙彼利埃 II 大学免疫遗传分子实验室（Laboratoire d'ImmunoGenetique Moleculaire）
• CINES（CNUSC）：法国蒙彼利埃的 d'Informatique National de l'Enseignement Supérieur 中心
• ICRF：英国伦敦的帝国癌症研究基金会（Imperial Cancer Research Fund）
• EBI：英国 Hinxton 的欧洲生物信息学研究所（European Bioinformatics Institute）
• IFG：德国科恩的遗传学研究所（Institute fur Genetik）
• BPRC：位于荷兰 RIJSWIJK 的灵长类生物医学研究中心（Biochemical Primate Research Centre）
• EUROGENETEC：位于比利时瑟兰（Seraing）

（5）IMGT 提供的服务和工具：

• 序列排列工具（如：DNAPLOT）
• 建模工具
• 作图数据分析工具
• 查询序列分类的工具
• 与其他的相关生物数据库有链接
• 通过网页界面直接提交数据

IMGT 有自己特有的编号方案。免疫效应分子的变异性越来越大，分类和分析这些分子需要一种与众不同的方法。IMGT 的编号方案解释分子结构框架（FR）、互补决定区（CDRs），如果分子有结构数据的话，描述其高可变区环的特征。

（6）搜索 IMGT 资料库特定序列时可采用的关键词类型：

• 受体类型（如：嵌合体、T 细胞受体、免疫球蛋白）
• 受体分类（如：TcRα、TcRβ、IgM、IgG）
• 侧链类型（如：TcRα 链、TcRβ 链、Ig 重链、Ig 轻链、Igκ 链）
• 区域类型（如：TcR 恒定区、Ig 恒定区、TcRα 恒定区、Igδ

恒定区）

• 描述性关键词（如： Fab、Fc、lambda5、transgene 等）

三、日本生物信息学服务器（GenomeNet-Japanese Bioinformatics Servers）

GenomeNet 是一家日本的数据库和计算机服务网络，专为分子和细胞生物学中基因组及其相关的研究领域服务。1991 年 9 月，在日本文部省的人类基因组计划开始实施时成立。GenomeNet 机构目前由京都大学化学研究所生物信息学中心（Bioinformatics Center，Institute for Chemical Research，Kyoto University）掌管。可通过 GenomeNet 服务器 http://www.genome.ad.jp／访问，并提供以下服务：

（1）DBGET/LinkDB 整和数据库检索系统（见图 3-8）

图 3-8 DBGET 数据库链接图，该图可以很方便地与其他数据库相链接，图中的连线说明了数据库相互之间的关系（http://www.genome.ad.jp/dbget/dbget.links.html）

• DBGET/LinkDB/KEGG 数据库链接图表

• DBGET/BLAST/FASTA 的 IDEAS 界面

（2）KEGG（Kyoto Encyclopedia of Genes and Genomes）：京都基因和基因组百科全书

- KEGG 内容表
- PATHWAY ：代谢途径和复合物
- GENES ：基因的注解
- SSDB：计算机分析的序列同源性
- LIGAND ：化学复合物和反应数据库
- EXPRESSION ：微阵列基因表达谱
- BRITE：蛋白和蛋白之间相互作用和关系
- BLAST/FASTA ：查找 GENES 和 GENOME 数据库的工具

（3）序列解译工具

- BLAST ：序列相似性查找
- FASTA ：序列相似性查找
- MOTIF ：序列中基序的查找
- CLUSTALW：多序列排列

（4）GenomeNet 的匿名 FTP 服务器，提供下载 KEGG 系统的路径。

下面就 4 个方面作一介绍：

1．GenomeNet 网站的链接

日本数据库系统不仅为美国、欧洲及日本的 DNA 和蛋白质数据库提供链接，而且确保这些数据库信息内容的质量。GenomeNet 开发的目的是解释序列信息以及为多种多样的生物学问题提供一系列最有用的分析工具，图 3-9 是其提供的工具。

GenomeNet 的工具包括：

- BLAST 和 FASTA ：均为序列相似性搜索程序。
- MOTIF ：由京都大学开发的序列特征搜索程序，搜索序列的特征不是线形排列，而与编码蛋白的结构特性直接相关。
- CLUSTALW：类似于 BLAST 和 FASTA 的多序列排列程

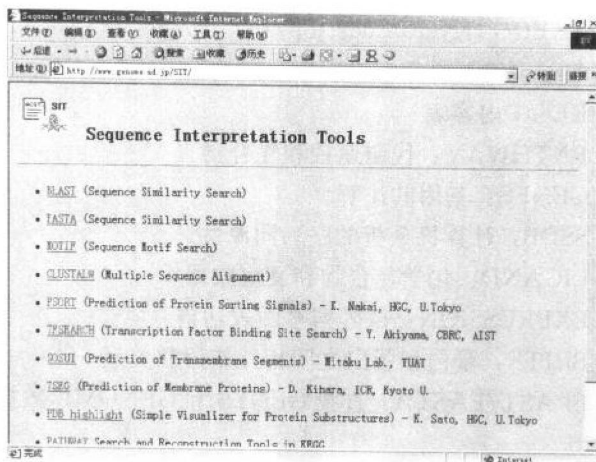


图 3-9 GenomeNet 提供的序列解释工具
(http://www.genome.ad.jp/SIT/)

序，仅限于成对比较，并提供其他一些工具搜索与蛋白结构相关的序列功能信息。

- TFSEARCH：识别转录因子结合位点。转录因子是通过直接结合到基因附近的 DNA 而控制基因活动的一组蛋白。
- GRAIL (Gene Recognition and Assembly Internet Link)：基因识别和装配互联网链接程序，用于识别在未知生物学意义的新近完成测序的 DNA 序列中的新基因。

因为基因含有一些结构用于调控蛋白的表达，所以某些短序列 DNA（10-50 个碱基对）的特殊结构和基序是功能基因的标志。而且，真核基因通常分成功能区和非功能区（编码和非编码区），称为外显子和内含子。只有外显子 DNA 序列能翻译成蛋白序列，因此，才有相应的蛋白序列和功能的解释或预测。尽管大部分数据库信息与序列有关，但蛋白结构信息也渐渐增多。从进化的角度来说，这很重要，因为蛋白结构较相应氨基酸和 DNA 序列更保守。为预测这些功能，可应用：

- PSORT 程序(prediction of protein sorting signals) 蛋白

分类信号预测

- SOSUI 程序(prediction of transmembrane segments)：跨膜片段预测
- PDB highlight ：是蛋白亚结构的简易观看程序，它可以模拟 Protein Data Bank 的蛋白信息用于比较已知蛋白结构的相似性。
- KEGG是有关生物代谢功能的一个工具。这是京都大学开发的关于代谢途径的搜索和重新构建的工具，下面将详细解释。

2.京都基因和基因组百科全书—KEGG(Kyoto Encyclopedia of Genes and Genomes)

不同生物的基因组测序进展顺利，小鼠基因组已有部分测序工作已完成；人类基因组目前已完成绝大部分的测序，计划将于2003年全部完成。在这些基因组计划中产生了大量的基因和基因序列的信息，下一步的工作就是诠释这些基因的功能。也就是利用实验和计算机的方法解码基因在生命体中实现功能的时间、地点和方式。KEGG 是 1995年 5 月在日本人类基因组计划的前提下启动的，KEGG 最初的目的是利用计算机的方法来分析目前已知的分子间相互关系的信息，包括代谢途径、调节途径和分子装配等方面的信息。KEGG 列出了所有生物中与代谢途径中的组分相关联的基因的目录。KEGG 包含很多已完成基因组测序的微生物的代谢途径信息，以及一些仍未完成基因组测序的生物体（如人和小鼠）的代谢途径信息。

KEGG 是 GenomeNet 数据库系统的一部分，与其他公共数据库通过 LIGAND 和 BRITE 搜索引擎相链接。

LIGAND 是一个化学数据库，用于搜索酶和代谢复合物。它由京都大学化学研究所维护，目前包含 13407个条目：3829 个相关酶(酶反应数据库)，9578 个相关代谢复合物（化学复合物数据库）。BRITE 为一生物分子相关信息传送和发布的数据库，也位于

京都大学化学研究所，含 278 个条目。

为了能直接搜索所关心的复合物、酶或基因，KEGG 主页上的“Search and compute with KEGG”链接将我们带入代谢图谱搜索工具、基因组图谱、着色工具、预测工具和序列相似性比较工具的界面。

只有知道需要查找的确切内容是什么时，这些链接才有用。因为这种搜索模式需要精确的条目号码，如酶术语 E. C. 2. 7. 1. 1 代表己糖激酶；化学复合物号码，如 C00417 代表顺式乌头酸盐。

对于更为普通的关键词搜索，或者只知道部分或完全酶、复合物或路径名称的话，则 LIGAND 搜索模式或代谢途径图谱和分子目录链接更为适合。后者在“pathway’,和“enzyme”选项中提供许多搜索类别。

3. KEGG 代谢数据库的应用

(1) 标准代谢路径的寻找：

图 3-10　赖氨酸生物合成标准代谢途径，图中显示了代谢途径中所有的酶，并且绘制了当前正在进行基因组计划的多种生物赖氨酸生物合成代谢途径（http://www. genome. ad. jp/kegg/pathway/map/map00300. html）.

自KEGG 内容表中点击路径分类下的“代谢路径”查看所有路径清单。为寻找“赖氨酸生物合成”路径链接，翻到路径名为“氨基酸代谢”一列，点击链接。并要查看赖氨酸生物合成的标准路径 MAP00300（见图 3-10）。

（2）寻找种属特异性路径图：

在标准路径图上，在“Go to”窗口中选择种属名（如 Mus musculus）并点击执行。那么，现在应该看到这些路径，并在窗口中会显示种属名（如 Mus musculus）。数据库现存的所有小鼠的酶都被标记为绿色，在 E.C.2.3.1. 酶组中有举例。人类相应图谱不显示单一的标记酶。对人和小鼠而言，L-赖氨酸是必需氨基酸。因为我们体内缺乏生物合成 L-赖氨酸的必需酶（见图 3-11）。

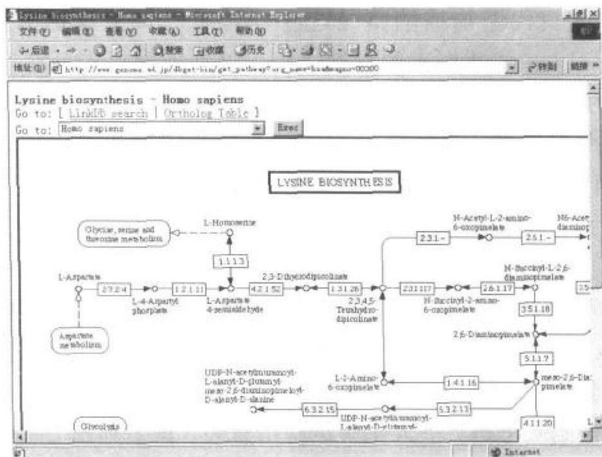


图 3-11　人类(Homo sapiens ）的赖氨酸生物合成代谢途径网页中有色彩的酶的编号表明存在这条途径，而没有色彩的酶编号表示不存在该条途径或还没有被描述，赖氨酸是人体必需氨基酸，人类缺乏合成赖氨酸的能力，所以在此图中几乎没有彩色的酶。

比较 E.coli 相应图谱，你会发现这种革兰 氏阴性菌能通过一系列的酶反应利用天门冬氨酸盐合成出 L-赖氨酸。

（3）从代谢路径图中找到化学结构或代谢信息：

注意：L-赖氨酸前体也可作为肽聚糖合成的底物，形成激活的前体分子。点击复合物下的小圆圈可以查看这种分子结构。还可于 KEGG 的 C05826 条目下查看信息页。

以上述及的问题给怎样寻找代谢信息提供了一个例子，即通过化学结构、分子式、KEGG 条目号、以及路径图查询。在路径图页面上点击底物名（或圆圈符号）是寻找底物相关化学信息和其公用路径最简易的方法。同样，点击 E.C. 号码框就可寻找酶，点击光滑边框符寻找交叉路径。例如，赖氨酸生物合成图有通往“赖氨酸降解”路径的链接。点击“赖氨酸降解”框符便可得到相应的代谢过程。选择 E.coli 路径，新路径图号码是 MAP00310。

（4）通过关键词搜索化学结构或代谢物信息：

为寻找代谢物或酶路径，内容列表提供了到达 KEGG 的 DBGET Ligand 数据库的直接链接。这种搜索模式可在 DBGET 搜索的“酶”目录下的“内容列表”上找到。点击“配基(Ligand)”链接上普通搜索模式，此时关键词就可应用。注意，DBGET 数据库无须准确的酶或复合物号码。寻找赖氨酸或 L-赖氨酸信息，键入“赖氨酸”敲回车便可。你可收到 96 个采样数的反馈清单，即反馈所有酶或复合物中包含“赖氨酸”的 KEGG 条目。清单中将有 45 个酶（ec：x.x.x.xx）和 51 个复合物（cpd：Cxxxxx），其中一个是“L-赖氨酸”(cpd：C00047）以及其他所有的衍生物。

点击 cpd 号码，进入化学结构信息表。该表列出 L-赖氨酸复合物查询号码（注意：不同于 D-赖氨酸)、公用名、分子式、结构、L-赖氨酸作为代谢物（包含 L-赖氨酸合成和降解的 5 个图、生物素代谢、生物碱生物合成及氨酰基-RNA 生物合成）的所有路径图，以及用 L-赖氨酸作为底物的所有已知酶。

4. 生物分子的一般信息资料

另一个有用的编排是分子目录条目，可更确切地说是“复合物(compound)分类”。这产生了代谢物根据其功能不同而分类的

目录。例如，碳水化合物、脂肪酸、磷脂和神经递质等。如想查找某一类分子的结构，诸如氨基酸或各种己糖，这个链接将给出最完善详尽的结果，可以用做查找结构信息的参照。例如，如果你对类固醇激素结构感兴趣，目录“脂类”将链接到含有 7 个胆固醇来源的类固醇激素名称及化学结构的页面。

点击“醛固酮”链接，链接到结构信息页，该页提供 C21 类固醇激素代谢（MAP00140）路径图链接。点击路径图链接，进入类固醇激素代谢的标准路径，醛固酮位置标记为红色圆圈，因为我们是从醛固酮开始研究的。选择 Homo sapiens，会显示出多个路径，而相应的细菌图 E. coli 显示出这种微生物缺乏合成类固醇激素的能力。

了解诸如 KEGG 数据库的局限性很重要。有时你想查看的酶没有被标记（如上述的醛固酮路径中）。该路径图显示所有已知的总结在标准路径图中的反应。种属特异性酶被标记成绿色。酶 EC1.14.15.5 是皮质酮 18-单加氧酶，功能是将皮质酮转化为醛固酮。从该酶的 E.C. 链接进入 GenBank 显示出一个小鼠的核苷酸序列，小鼠的 CYP11B2 基因第 9 外显子负责醛固酮合成。人类单加氧酶的同源基因同样存在，但还未见到报道。

以上描述的工具和数据库可以通过 Japanese Bioinformatics Servers（日本生物信息学服务器）的 http://www.genome.ad.jp 上获得。

第二节　数据库开发工具

一、序列相似性搜索工具

序列相似性搜索工具（sequence similarity search tools）是指查找序列之间的同源序列的工具，用来明确序列之间相似性的大小。本节主要讲述 BLAST 和 FASTA，这是互联网上最流行、界面友好的两大序列相似性搜索工具。 BLAST 服务器由美国的

NCBI 支持，而 FASTA 则由英国的 EBI 负责维护。BLAST 在 EBI 的镜像站上提供了用户应用 BLAST 或 FASTA 的选择权，并且提供一些其他有用的搜索程序。但 NCBI 用户仅限于使用 BLAST 服务器。非常有效，适合多种搜索任务。本章将进一步讨论这一（见图 3-12）。其他序列相似性搜索工具上的附加信息可从 EBI 和 NCBI 网站上获得。

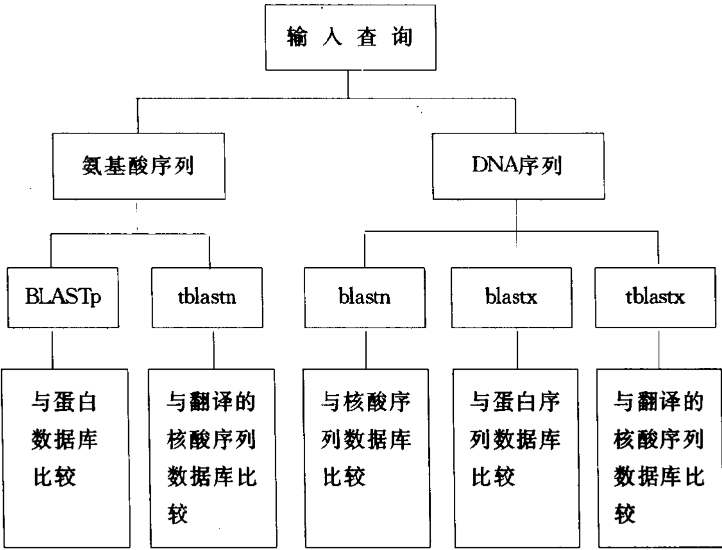


图 3-12　BLAST程序一览图

序列相似性搜索是通过序列排列的方法实现的。大体了解序列排列（sequence alignment）对理解 BLAST 或者其他序列相似性搜索工具是必须的。下面简要介绍一下这种大多数序列相似性搜索工具的基础——序列排列。

1. 序列排列：

多数情况下，序列排列用于发现潜在的同源性，继而预测寻找序列的潜在功能或帮助模拟其三维结构。序列排列工具分为整

体或局部排列工具（global or local alignment tools）两类。

（1）整体排列工具（global alignment tools）：

整体排列工具是对特定序列全长最好的总体排列。两条序列间引入间隙（gap）允许全长序列的总体排列。应用整体排列工具的主要优点是可使具有高度相似性的序列得到最优化的处理。依据与已知三维结构的序列同源性进行结构模拟预测时，该工具是非常有用的。

（2）局部排列工具（local alignment tools）：

局部排列工具是在特定序列的亚区或局部区域寻找优化排列。应用局部排列工具的优点是它对局部呈现相似性区域的序列最为适合。局部排列搜索工具用于寻找序列基序、结构域和同一序列内的其他类型的重复序列。在给定的数据库内，寻找相似的序列也很有用。总之，局部排列工具特别适合识别高度相似的较短区域记分的片段。通常这些区域内片段可以用于发现全长序列的相似性。

2. 两种序列排列工具的记分方案：

包括 BLAST 和 FASTA 两种序列排列工具在内，所有序列比较算法都依赖于某一种记分方案。大多数采用记分矩阵给每一个排列记分。排列分值是赋予每一个配对的氨基酸或核苷酸对的较小分值的总和。区分记分矩阵的标准依赖于它所依靠的记分类型。大多数矩阵依赖以下记分方案中的一种。

（1）依据“同一性”的记分方案：

在这个记分方案中，配对的相同残基或核苷酸记为正分，而不相同配对记零分。一般地，赋予相同配对的正分为 1。总体同一性分值转换成同一性百分比。

其优点是具有简单和非启发性，对于有高度序列相似性的序列很好用。

但缺点是这种记分方案总体上不如考虑了外部知识的记分方案。这主要由于非同一性配对也有不相同之处。例如，从生物角

度看，丙氨酸／缬氨酸配对较丙氨酸／天冬氨酸盐对更能接受。在这个例子中，不同之处是所涉及残基的相对疏水的性质。因此，“同一性”的记分方案在监测低序列相似性的序列或序列区域时不很有效。因而，一种考虑了额外步骤的非同一性配对记分方案较单纯同一性记分更具生物学意义。还有，从排列所得的同一性百分比报告不总是与所呈现的同源程度相一致。这主要因为长度影响同一性百分值。

（2）依据“化学相似性”的记分方案：

这是为了克服“同一性”记分方案缺点的一种基本尝试。这种方法考虑了残基对的化学结构特性。 McLachlan 和 Feng 的记分方案都考虑了氨基酸的特性，如极性、电荷以及结构特征等。

其优点是在某种程度上，它与氨基酸水平的蛋白结构的真正选择性压力一致。事实上，有些氨基酸的突变与其它各种突变相比，对于蛋白功能具有较大的破坏性。一般来说，这些突变都会引起相应氨基酸特性的剧烈改变。极性残基转变为非极性残基，或相反的情况，对于改变蛋白的结构和功能较具有相似特性的残基突变更为有效。

其缺点是在自然状态下观察的突变不是总能从简单的记分方案中得到解释，虽然这种记分方案体现了对自然现象的基本理解。自然界中某些进化突变仍需进一步研究。

（3）依据“遗传编码”的记分方案：

这种方法考虑了在基因组水平上导致由一个氨基酸转换为另一个氨基酸的碱基变化的最小值。其优点是基于分子生物学的原理。缺点是偶然性因素可能会影响该方案的可靠性。改变了的残基仍可能在较大程度上保持相似性，而较低的碱基改变率并不总是与此一致。

（4）依据“观察突变”的记分方案：

该方法以在排列序列中观察到的突变频率为依据。优点是以自然状态下真实发生的情况为基础，可以最大程度地减少某种臆

断。缺点是该记分矩阵基于在一套排列序列中发现的突变频率，由于最初排列时需要人为地干预，有可能会改变观察的真正突变频度。通过肉眼的序列排列亦可能会产生配对错误，并最终导致非自然突变频率的发生。一般地说，依据“观察突变”的记分方案与以上几种方案比较，它较好地体现了自然变化的过程。

3. 序列排列的用途：

- 进化：序列间的同源性很高往往暗示着相互之间有较近的进化关系。
- 结构预测：未知结构的蛋白序列与蛋白结构已知的序列排列可以预测那些未知的三维结构，但预测的结构仍需实验鉴定。这是基于这样的假设：即在相关的蛋白中，序列同源性和结构相似性有直接关系。
- 序列基序（motif）鉴定：局部序列排列可以鉴定出蛋白和核苷酸的潜在序列基序和功能特征。
- 功能预测：蛋白间的高度序列相似性通常暗示所分析的同源序列功能可能相同。

4. 大多数蛋白序列算法的基本概念：

这些算法基于 210 个可能的氨基酸配对,由 20×20 的记分矩阵加以描述。210 是 20个配对的和 190 个不配对的氨基酸对的总和。在给定的字符表中，字符总的可能对数由公式（n-1）i 表示（n 代表氨基酸字符的数目）。因此，有 20 个氨基酸符号的蛋白质用（20-1）i 表示，与 210 个可能的氨基酸对相对应。正如前面所讨论的，在记分矩阵中，相同的氨基酸对（如亮氨酸和亮氨酸）被赋予最高分，接下来是某种程度相似的氨基酸对（如亮氨酸和异亮氨酸），最后才是不相似的氨基酸对（如亮氨酸和精氨酸）。

5. NCBI 的同源搜索基本工具-BLAST（Basic Local Alignment Search Tool）

BLAST 可以搜索所有可获得的主要序列数据库（如 SWISS-PROT，PDB 等）。标准 BLAST 的默认运行数据库是 nr（non-

redundant）数据库，nr 数据库由 NCBI 维护。因其缺乏同一种属的冗余序列，故而加速了 BLAST 对输入文件的分析。虽然 nr 数据库是 BLAST 运行的默认数据库，用户仍可选择其想要查询的其他数据库。例如，如果用户想用结构已知的同源蛋白用来模拟结构尚需确定的蛋白序列的话，PDB (Protein Data Bank)数据库将是最合理的选择之一。

（1）BLAST 中的记分矩阵

BLAST 的统计理论由 Samuel Karlin 和 Steven Altschul 创立。所有的 BLAST 程序应用替代记分矩阵（substitution scoring matrix）。排列过程中的扫描相和扩展相都应用替代矩阵。该矩阵用于给配对记分。已知替代矩阵能很大程度地加强排列过程的敏感性。这对于 BLAST 试图发现序列的相似性部分或片段是至关重要的。替代矩阵是一种记分方法，用于一个氨基酸残基或核苷酸与另一个残基或核苷酸的排列。替代矩阵的首次应用是用进化的角度比较蛋白的序列，由已故的 Margaret Dayhoff 和她的同事共同开发。这些矩阵来源于近似序列的整体排列，同时也用于外推相似性较弱的或进化距离较远序列的其他矩阵。这些矩阵专指 Dayhoff，MDM 和 PAM 系列矩阵。这些矩阵的相关数字（如 PAM40，PAM100 等）是与各个序列间的相应进化距离相一致的。较小的数字表示进化距离较小的序列，而较大的数字代表较远的进化距离。PAM 系列矩阵的主要缺陷是其基于不正确的假设上：即相关序列间的选择压力与不大相关序列间的选择压力是一样的。由 Steve Henikoff 及其同事开发的 BLOSUM 矩阵，与依赖相关序列的整体排列的 PAM 系列矩阵不同，它来源于相关序列的局部排列。BLOSUM 矩阵不依靠原先计算的不太相关的序列矩阵来外推较为相关的序列。这种方法的所有矩阵都直接由计算得出。与 PAM 系列矩阵相比，伴随 BLOSUM 矩阵的数字（如 BLOSUM62)指的是用于构建矩阵的最小同一性百分比。因此，较小的数字与代表进化距离较大的间隔相一致。

（2）可以使用的矩阵

PAM 系列一般适合整体相似性的搜索，而 BLOSUM 系列则能较好地寻找区域或局部的相似性序列。两个系列都有其优缺点，现行的办法是合并使用两种方法，使其优势互补。这样一个结合的矩阵能够在相似性搜索工作中提高操作的水平。

BLAST 程序的设计是为加快速度，同时最大限度地增加对序列距离关系的敏感性。这就使得该程序以时间高效性的方式寻找同源性最近的序列。BLAST 程序应用启发式算法确定局部排列。与寻找整体排列的算法相比，BLAST 的局部排列搜索是寻找序列相似性的孤立区域的。BLAST 服务器支持多种分析程序，既可通过网页界面获得，也可安装在局域网上，以加快分析的步骤。Standard BLAST 是最初的 BLAST 程序，仅能在 NCBI 数据库网络中搜索相似性序列。

（3）BLAST 基本搜索（basic BLAST search）的局限性

基本 BLAST 程序在其排列中不允许有缺口的出现。从理论上讲，缺口的出现将减低搜索的敏感性。但是，输出文件所显示出的多个区域的排列可以用作预测查询序列和数据库序列间的缺口（gap）

（4）不同的 BLAST 程序及其用处

• BLASTp ：该程序允许用户在蛋白数据库中搜索所需要的蛋白序列。可用于在查询序列数据库中寻找与某一已知蛋白可能同源的所有序列。

• BLASTx ：该程序允许用户在蛋白数据库中搜索翻译的核苷酸序列。被查询的核苷酸序列先被翻译成 6 个可能的阅读框。翻译后的核苷酸序列通过与蛋白序列库进行比较，发现其中可能的同源性蛋白。也可用于寻找核苷酸测序中的错误。在这种情况下，该程序尤其有用。BLASTx 输出文件的信息也可以帮助鉴定特定核苷酸序列中尚不清楚的核苷酸。

• BLASTn ：该程序允许用户在核苷酸数据库中搜索与查询

序列同源的核苷酸序列。一个新近测序的核苷酸可以与其同源体进行比较，以对该序列进行鉴定或发现是否有可能有其他序列污染。

· tBLASTn ：允许用户以蛋白查询序列在特定的核苷酸数据库中搜索翻译的核苷酸序列。在特定的核苷酸数据库中核苷酸序列一开始被翻译成 6 个可能的阅读框，然后与蛋白查询序列进行比较。通过将蛋白查询序列与特定的核苷酸数据库中翻译的核苷酸同源体进行比较来发现蛋白测序错误时，该程序特别有用。tBLASTn 输出文件中的信息还有助于澄清特定查询序列中不明的氨基酸残基。就 6 个阅读框翻译比较方法而言， tBLASTn 与 BLASTx 相似，但它不是使用核苷酸查询序列（用于 BLASTx）去查询，而用蛋白查询序列。由于该程序需要将特定核苷酸序列数据库中所有序列都翻译成 6 个可能的读框，所以运行时间非常长。

· tBLASTx：该程序首先将查询的核苷酸序列翻译成 6 个可能的读框，然后将核苷酸序列数据库中所有序列也翻译为 6 个可能的读框，最后将查询序列的翻译结果与核苷酸序列数据库的翻译结果进行同源性比较，以发现同源序列。 tBLASTx 与 BLASTx 和 tBLASTn 的程序相似，是 BLASTx 搜索的补充。

(5) 新 BLAST 程序：

新的 BLAST 程序称为 BLAST2.0。Gapped BLAST 和 PSI-BLAST 是 BLAST2.0 服务器支持的两个应用程序。新的 BLAST2.0 服务器已重新设计，以优化速度和灵敏性，并新增了支持 Gapped BLAST 和 PSI-BLAST 应用程序的能力。

(6) Gapped BLAST：

Gapped BLAST 的算法允许在序列排列中引入缺口（gap）以输出 BLAST 文件。缺口是序列中缺失和插入的部分。这种方法避免了相似序列区被分割成片段。该算法的探索性使其输出分值能反映相关序列的生物学关系。一般地说，这反映了保存完好的序列活动区和结合区的情况。因此，缺口的引入避免了这些区域被

离散成无意义的序列片段。

（7）PSI-BLAST:

PSI-BLAST 代表位点特异性迭代的 BLAST。PSI-BLAST 开始执行 Gapped BLAST ,用输出序列作为它自己的输入文件。这样，PSI-BLAST 便构建了一个位点特异性的记分矩阵，该矩阵取代了原始查询序列，在接下来的几个重复数据库搜索操作中寻找用户感兴趣的主题。主题查找增加了同源序列查询的灵敏性。

有些 BLAST 工具可以安装在局域网微机中。这是 BLAST 网络客户机的程序软件。局域网的网络客户机程序软件与远程 BLAST 服务器（NCBI）的 BLAST2 和 Power BLAST 间的信息交流是 BLAST 提供的基本的网络服务。

（8）BLAST2:

BLAST2 是 BLAST 的标准服务，用于比较两个序列之间的同源性,提供 HTML 格式的输出文件。其滤过能力能使用户找到低复杂区序列。

（9）PowerBLAST 用途：

这是一个网络 BLAST 的客户机程序，实施大范围的基因组信息的分析任务。该程序以及其他网络客户机软件可在 BLAST 的网络目录下，经 FTP 自 NCBI 主页检索。

进入 BLAST 服务器的方法有几种。最便捷的方法是通过网址（http://www.ncbi.nlm.nih.gov） BLAST 运行中的网页界面极其友好。以下是用户成功运行 BLAST 必须的一般步骤：

·所关心的查询序列必须有正确格式（如，FASTA 格式,只包含序列的一种格式）。如果查询序列是从 NCBI 的 Entrez 检索所得的话，最简易的办法是自 Entrez 复制该序列的 FASTA 格式，粘贴到 BLAST 界面中。

·接着，将编成适当格式的序列粘贴到 BLAST 网页的“序列输入”框中。

·依据分析序列的类型，选择适当的 BLAST 程序。（如，蛋

白序列选 BLASTp，DNA 或 RNA 选 BLASTn，等等），新的 BLAST 网页界面要先选择要应用的程序，然后再粘贴序列。

· 最后，选择适当的数据库。BLAST 的默认数据库是 NCBI 的 nr 数据库。nr 数据库将搜索现有所有的非冗余序列。例如，如用户只想查询结构已知的同源性序列，那么，搜索已知分子结构的特异性数据库是较为明智的。因此，用户可以选择 PDB 作为首选数据库。然后点击"Submit"链接将序列发送到 BLAST 服务器。搜索结果既可通过电子邮件获得，也可在 BLAST 网页界面查看。分析多个序列文件时，电子邮件途径较为理想。它可使用户高效率地分析感兴趣的序列，并且能分析后来的部分结果（BLAST 图片可以保存为 GIF 文件)。

如前面叙述的那样，BLAST 还可经 BLAST 网络客户程序获得。这样的话，用户首先必须通过 FTP（ftp://ncbi.nlm.nih.gov)安装合适的 BLAST 网络客户程序软件。还可通过 NCBI 的电子邮件服务器（blast@ncbi.nlm.nih.gov)完成 BLAST 搜索。这主要适用于不便上网的人。同样，查询序列必须有适当的格式，以便 BLAST 能完成相应的操作。运行 BLAST 的另一个方法是，在局域网微机上安装完全可执行版本，搜索用户局部数据库。BLAST 的这个版本可在 BLAST 的"executables"路径下找到，可经 FTP（ftp://ncbi.nlm.nih.gov ）获得。可以获得 BLAST 用于 IRIX6.2、Solaris2.5、DEC OSF1 及 Win32 操作系统的版本。BLAST 查询结果中检索的序列与 NCBI 的 Entrez 以及 PubMed 服务器有直接或间接的链接，从而可以得到查询结果的序列和相关文献。

（10）BLAST 输出文件中期望值（expect，E 值）的意义：

在特定的数据库中，期望偶然匹配的几率大小称为 E 值。因此，得到较低 E 值的查询结果有意义。E 值为零意味着这一特殊的查询结果被随机匹配的可能性是零。这样，一个结果的 E 值表明在特定的数据库中发现相似序列匹配的可能性。 E 值是特定匹

配中基本的随机噪声。随记分值（score，S 值）的增加，E 值呈指数性减少，即随机噪声减低，表明序列同源性较高。可以增加 E 值以发现统计学意义较小的结果，对于统计学意义较小的短肽或短核苷酸序列，加大 E 值可能会得到查询结果。

6. EBI 的同源搜索工具——FASTA

FASTA 是 EBI 提供的同源搜索工具，目前 EBI 网页提供的最新版本是 FASTA3（http://www2.ebi.ac.uk/fasta3/），FASTA3 可以接受多种序列格式的查询，如：FASTA、GCG、EMBL、Genbank、NBRF 和 Phylip 等。对于短序列的查询（1-6 个核苷酸），FASTA3 的输出结果没有 BLAST 多，但其结果的相关性更高。FASTA3 提供多种应用程序的选择：

- fasta3：查询序列与一个 DNA 或蛋白质数据库同源性比较。
- fastx/y3 用于一个翻译的 DNA 序列 6 个可能的读框与蛋白质数据库同源性比较。
- tfastx/y3：用于一个蛋白序列与翻译的 DNA 序列数据库中所有序列比较。
- fasts3：用于连接肽与蛋白质数据库比较。
- fastf3：用于混合肽与蛋白质数据库比较。

FASTA3 同源性比较中可以选择的数据库：

- swall：SWALL 非冗余蛋白质序列数据库（Swissprot＋Trembl＋TremblNew）
- swissprot：SWISS-PROT 蛋白质序列数据库
- swnew：SWISS-PROT 的更新数据库
- sptrembl：SPTREMBL（TrEMBL）数据库
- remtrembl：REMTREMBL（TrEMBL 中未处理的条目）
- PDB：Brookhaven 的蛋白质数据库
- ENSEMBL：ENSEMBL 编码区数据库
- Euro Pat：欧洲专利局(European Patent Office)蛋白序列

专利数据库

- Japan Pat：日本专利局(Japanese Patent Office)蛋白序列专利数据库
- USPTO Pat：美国专利商标局（United States Patent and Trademark Office 蛋白序列专利数据库
- EMBL：EMBL 核苷酸序列数据库
- EFUN：EMBL 真菌序列数据库
- EINV：EMBL 无脊椎动物序列数据库
- EHUM：EMBL 人类序列数据库
- EMAM：EMBL 哺乳动物序列数据库
- EORG：EMBL 细胞器序列数据库
- EPHG：EMBL 噬菌体序列数据库
- EPLN：EMBL 植物序列数据库
- EPRO：EMBL 原核生物序列数据库
- EROD：EMBL 啮齿动物序列数据库
- ESTS：EMBL 序列标签位点数据库
- ESYN：EMBL 合成序列数据库
- EUNA：EMBL 未分类序列数据库
- EVRL：EMBL 病毒序列数据库
- EVRT：EMBL 脊椎动物序列数据库
- EEST：EMBL 表达序列标签数据库
- EGSS：EMBL 基因组探察序列数据库
- EHTG：EMBL 高通量基因组序列数据库
- EMNEW：EMBL 更新序列数据库
- EMALL：EMBL＋EMBL 更新序列数据库
- IMGT：IMGT 免疫遗传标记数据库
- HGBASE：欧洲单核苷酸多态性数据库

FASTA3 有自己的一套独特记分方法，这里就不在赘述。用户可以根据自己的需要选择不同的程序和数据库达到自己的目

的。

7. 数据库序列搜索概述

序列搜索是指将一个已知序列与数据库中的序列进行比较，以发现数据库中与已知序列同源的序列。

（1） 目的：

• 寻找同源性序列，推断查询序列的特性。

• 鉴定已知三维结构的同源性序列，预测靶序列的三维结构，推断其功能特性。

（2） 可能的问题：

在特定的数据库中，能否区分是真正的同源性序列还是碰巧发现的序列，仍是一个问题。这些不能辨别的结果必须经过进一步地检测，以了解其与查询序列的真正关系。

（3） 序列数据库搜索的方法：

• 需要一个查询序列，即需要分析的靶序列。查询序列可以是一个新近测定的序列，性质有待鉴定；也可以是特性已知的序列。数据库搜索可帮助确定新近测定的序列性质，或者帮助某一已知查询序列条目发现其可能的序列同源体。

• 选择适当的服务器。服务器必须是可靠的、定期更新的以及有影响的。这些特点一般让人想起政府或政府资助的生物信息服务器，如：NCBI。NCBI 是几个公共领域数据库和搜索工具的集合，易于通过互联网获得，且与大多数的网页浏览器兼容。当然，EBI 也是很好的选择。

• 在特定的服务器中，选择合适的程序或程序组。如果选定了 NCBI 服务器，并需要一个执行简单的序列相似性搜索的工具的话，那么，BLAST 程序便很合适。同样也可以选择 EBI 的 FASTA 程序。

• 选择一个合适的 BLAST 程序用于简单的序列相似性搜索，如果所查询的是蛋白序列， BLASTp 是合适的工具。如果要查询 DNA 或 RNA 序列，则必须应用 BLASTn 程序。它们仅是

BLAST 服务器多个数据库中的两个程序。其他的 BLAST 程序（如 tBLASTn，tBLASTx）可用于寻找被查询序列的同源性序列，也可执行更高级任务。例如， BLASTx 程序可用于发现基因中的潜在编码区；而其他的 BLAST 程序可用于查对新近确定的序列中可能存在的测序错误。

· 选择适当的数据库。有两种方法：其一，搜索包含所有相关序列的数据库，这是一个包含了所有提交条目的非冗余数据库。这种方法能使用户在所有序列条目中进行搜索。其二，搜索特定的数据库。在这种情形下，用户只须关心某一特定的数据库。例如，如果用户想寻找已知三维结构的同源性序列， PDB（Protein Data Bank）数据库则是最合理的选择，因为其所有序列条目的三维结构都是已知的。

· 选择适当的滤器（filter）。为方便用户，BLAST 在每一个程序中插入了一套滤器选择项。滤器选项可以排除低复杂区序列。由于序列的重复属性，在搜索内的假阳性结果或随机结果的概率增加，最终使结果模糊不清。我们推荐在特定的搜索中使用滤器，以减小假阳性的数量。但滤器选项有可能从结果中排除掉真阳性。低复杂区真阳性结果则可能从输出文件中排除。所以滤器选项可能会降低搜索的灵敏性。如何才能最大限度地提高搜索的灵敏性，同时又可减少假阳性的发生？这可通过对同一查询序列实施两次不同的搜索而实现。在一次搜索中，应用滤器，减少假阳性结果；而另一次不用滤器，以增加灵敏性。将两次搜索结果的输出文件进行比较，找出用滤器排除掉的可能的真阳性结果。

· 阅读、理解及分析输出文件。为从搜索查询结果中得出可能的假设，用户需要熟悉输出文件中的术语。输出文件的关键要素是每一个查询结果的分值和数据库给予每个结果的序列号。每个查询结果的分值暗示其与查询序列的同源性。在 BLAST 输出文件中，同源性也与赋予每个结果的期望值（E 值）相关。E 值是随机或偶然采集序列的概率，越接近零，在特定的数据库中，被

随机采集的可能性就越小。

（4）应用网络数据库相对于局域数据库的优点：

· 网络数据库是定期更新的。 NCBI 和欧洲的 EBI 及日本的 DDBJ 通力合作，使他们的数据库每日都在更新。这种每日的更新给其用户提供了可靠和非冗余的资源。

· 对局域数据库的维护决非小事。在大多数情况下，超出了一般用户的能力。应用和维护个人数据库费时，而且昂贵。这些障碍提高了 NCBI、EBI 以及 DDBJ 等公共领域网络数据库的价值。

· 网络数据库给他们的用户提供了适当的搜索工具。 NCBI 提供给用户 BLAST 服务器，EBI 和 DDBJ 也一样。通过公共领域服务器提供的搜索工具也能定期更新，这也有利于用户的使用。

（5）应用网络数据库而不应用局域数据库的缺点：

· 网络瘫痪时，局域数据库则易于登陆。

· 用户受限于网络数据库所提供的搜索工具。网络服务器所采用的扫描方法并不总是最高效的。通过局域服务器采用局部扫描的方法可能对某一特定的搜索更为合适。

本节所描述的 BLAST 程序可通过 BLAST 服务器 www.ncbi.nlm.nih.gov（NCBI 主页）获得。

二、特征识别工具和数据库

Prosite 是最广泛应用的数据库之一，包含生物基序（motif）和识别标志。Prosite 是一个在许多蛋白中发现的功能位点和序列特征的集合。

1. Prosite 数据库储存的信息及对用户的作用

Prosite 收集和拥有许多特征性结合位点和基序。在多数情况下，Prosite 的条目与其他适当的网站相互联系，并相互参考。例如，Prosite 详细记录了钙结合位点 EF-hand 的识别标志，该条目特征是由 SWISS-PROT 文件详细描述的。这些条目一般都与 SWISS-PROT 及其他相关数据库相联系。

· Prosite 文件包括：含有所关心的配对序列的基序或识别标志。该文件同时通知用户假阳性、假阴性的可能性以及有疑问的配对序列。假阳性序列是一类偶然性的识别标志或基序。假阳性序列一般缺乏所关心序列基序的功能特性。假阴性序列是与查询序列具有真正相同功能的结果，但缺乏特异性识别信号的一类序列。可疑序列是与查询序列有共同的基序特征但功能意义尚不能经实验证实的一组序列。通过实验可以将这些可疑序列进行假阳性或真阳性分类。这种信息类型给用户提供强大的工具，提高其工作效率。

· Prosite 拒绝冗余信息。为了减少冗余基序，已经详细研究了特征性识别信号。

· Prosite 有特征配对的搜索工具。可应用 PROMOT 搜索工具在 Prosite 数据库中找寻配对序列，也可用它在一些给定的特征中与感兴趣的序列配对。Prosearch 是另一个搜索工具，可以用特定的序列特征或识别信号来查找 SWISS-PROT 及 Tremble 的数据库。通过 Prosearch ，可以在 SWISS-PROT 和 TREMBL 所有序列条目中有效地发现新的序列识别信号和特征。

2. Prosite 文件资料的提供方式：

每个序列特征文件以“ . doc ”文件出现，而实际的序列特征则出现在另一个分开的文件中，标记为“ dat ”文件。dat 文件包含了特征扫描程序和其他序列特征编辑程序所需要的有关信息。

3. 识别信号的含义及阅读和构建的方法：

为了更好地理解 Prosite 数据库中用于每个识别信号中的符号，以钙结合位点 EF-hand 序列基序为例，表示 Prosite 以钙结合位点 EF-hand 的识别信号是：D-X-[DNS]-{DENSTG}-[DNQGHRK]-GP}-[LIVMC]-[DENQSTAGC]-X (2)-[DE]-[LIVMFYW]

说明：

1. 连字符用来分离序列基序中的每个位置。

2. [ ]：每个括弧中的残基代表序列基序中某一特殊位置允许出现的残基。例如，在［DNS］中，在其特定位置允许的残基是天门冬氨酸、天冬酰胺和丝氨酸。

3. { }：大括号中的符号代表序列基序中特定位置不允许出现的残基。换句话说，该特定位置允许出现其他残基。

4. X：表示二十个氨基酸中的任何一个。

5. (n)：代表某特定残基或氨基酸 X 的重复数。例如，X(2)代表-X-X-。

6. (n，m)：代表 n 和 m 间一段序列的重复长度。例如， A(2，5)意味着在序列基序中的一个特定位置上，可能出现连续 2、3、4 或 5 个丙氨酸。

参考文献：

1. Woodsmall RM, Benson DA. Information resources at the National Center for Biotechnology Information. Bull Med Libr Assoc, 1993, 81 (3): 282-4
2. Emmert DB, et al. The European Bioinformatic Institute (EBI) databases. Nucleic Acids Res, 1994, 22 (17): 3445-9
3. Barker WC, et al. The PIR-International Protein Sequence Database. Nucleic Acids Res, 1998, 26 (1): 27-32
4. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. Nucleic Acids Res, 1998, 26 (1): 38-42
5. Sussman JL, et al. Protein Data Bank (PDB): database of three dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr, 1998, 54 (1): 1078-84
6. Benson D, Lipman DJ, Ostell J. GenBank. Nucleic Acids

Res, 1993, 21 (13): 2963-5

7. Boguski MS, Lowe TM, Tolstoshev CM. dbEST-database for " expressed sequence tags" [letter]. Nat Genet, 1993, 4 (4): 332-3
8. Moyzis RK, et al. The distribution of interspersed repetitive DNA sequences in the human genome. Genomics, 1989, 4 (3): 273-89
9. VanBogelen RA, et al. The gene-protein database of Escherichia coli: edition 5. Electrophoresis, 1992, 13 (12): 1014-54
10. Martin AC. Accessing the Kabat antibody sequence database by computer. Proteins, 1996, 25 (1): 130-3
11. Payne WE, Garrels JI. Yeast protein database (YPD): a database for the complete proteome of Saccharomyces cerevisiae. Nucleic Acids Res, 1997, 25 (1): 57-62
12. Cavin PR, Junier T, Bucher P. The Eukaryotic Promoter Database EPD. Nucleic Acids Res, 1998, 26 (1): 353-7
13. Altschul SF, et al. Basic local alignment search tool. J Mol Biol, 1990, 215 (3): 403-10
14. McEntyre J. Linking up with Entrez. Trends Genet, 1998, 14 (1): 39-40
15. Rashbass J. Online Mendelian Inheritance in Man. Trends Genet, 1995, 11 (7): 291-2
16. Ohkawa H, Ostell J, Bryant S. MMDB: an ASN. 1 specification for macromolecular structure. Ismb, 1995, 3: 259-67
17. Hogue CW. Cn3-D: a new generation of three-dimensional molecular structure viewer. Trends Biochem Sci, 1997, 22 (8): 314-6

18. 王哲、黄高升。NCBI 的数据库资源及应用。生命科学，2002，14（1）：59-62
19. 张玲。网络生物资源及其利用。国外医学分子生物学分册，1999，6（21）：321-325
20. Pearson WR. Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol，1994，25：365-89
21. Brenner SE. BLAST，Blitz，BLOCKS and BEAUTY：sequence comparison on the net. Trends Genet，1995，11（8）：330-1
22. Stoesser G，et al. The EMBL nucleotide sequence database. Nucleic Acids Res，1998，26（1）：8-15
23. Bairoch A，Apweiler R. The SWISS-PROT protéin sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res，1999，27（1）：49-54
24. Rodriguez-Tome P，Lijinzaad P. The radiation hybrid database. Nucleic Acids Res，1997，25（1）：81-4
25. Rodriguez-Tome P. Searching the dbEST database. Methods Mol Biol，1997，69：269-83
26. Lefranc MP，et al. IMGT，the International Im Muno Gene Tics database. Nucleic Acids Res，1998，26（1）：297-303
27. Berman H. M，Zardecki C，Westbrook J. The nucleic acid database：a resource for nucleic acid science. Acta Crystallogr D Biol Crystallogr，1998，54（1）：1095-104
28. FlyBase：a Drosphila databse. Flybase Consortium. Nucleic Acids Res，1998，26（1）：85-8
29. Attimonelli M，et al. MitBASE：a comprehensive and integrated mitochondrial DNA database. Nucleic Acids Res 1999，27（1）：128-33

30. Moszer I, Glaser P, Danchin A. SubtiList: a relational database for the Bacillus subtilis genome. Microbiology, 1995, 141 (Pt 2): 261-8
31. Barker WC, et al. Protein sequence database of the protein identification resource (PIR). Protein Seq Data Anal, 1987, 1 (1): 43-98
32. Bairoch A, Bucher P, Hofmann K. The PROSITE database, its status in 1997. Nucleic Acids Res, 1997, 25 (1): 217-21
33. Boehnke M, Lange K, Cox DR. Statistical methods for multipoint radiation hybrid mapping. Am J Hum Genet, 1991, 49 (6): 1174-88
34. Matise TC, Perlin M, Chakravarti A. Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map [published erratum appears in Nat Genet, 1994, Jun; 7 (2): 215]. Nat Genet, 1994, 6 (4): 384-90
35. Lefranc MP, et al. ligm-db/imgt: an integrated database of Ig and TcR, part of the immunogenetics database. Ann N Y Acad Sci, 1995, 764: 47-9
36. Newell WR, Trowsdale J, Beck S. MHCDB-database of the human MHC. Immunogenetics, 1994, 40 (2): 109-15
37. Kanehisa M. Linking databases and organisms: GenomeNet resources in Japan. Trends Biochem Sci, 1997, 22(11):442-4
38. Fujibuchi W, et al. DBGET/LinkDB: an integrated database retrieval system. Pac Symp Biocomput, 1998: 683-94
39. Altschul SF, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 1999, 27 (1): 29-34
40. Pearson WR. Using the FASTA program to search protein

and DNA sequence databases. Methods Mol Biol, 1994, 24: 307-31

41. Thompson JD, Higgins DG, Gibson TJ. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 1994, 22 (22): 4673-80
42. Roberts L. GRAIL seeks out genes buried in DNA sequence [news]. Science, 1991, 254 (5033): 805
43. Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci, 1999, 24 (1): 34-6
44. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure predication system for membrane proteins. Bioinformatics, 1998, 14 (4): 378-9
45. Goto S, Nishioka T, Kanehisa M. LIGAND dataase for enzymes, compounds and reactions. Nucleic Acids Res, 1999, 27 (1): 377-9
46. Bairoch A. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res, 1991, 19 (Suppl): 2241-5
47. Bairoch A, Boeckmann A. The SWISS-PROT protein sequence data bank. Nucleic Acids Res, 1991, 19 (Suppl): 2247-9
48. Sternberg MJ. PROMOT: a FORTRAN program to scan protein sequences against a library of known motifs. Comput Appl Biosci, 1991, 7 (2): 257-60
49. Kolakowski LF, Leunissen JA, Smith JE. ProSearch: fast searching of protein sequences with regular expression patterns related to protein structure and function.

Biotechniques, 1992, 13 (6): 919-21

50. Huang X. On global sequence alignmet. Comput Appl Biosci, 1994, 10 (3): 227-35
51. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol, 1996, 266: 460-80
52. Feng DF, Johnson MS, Doolittle RF. Aligning amino acid sequences: comparison of commonly used methods. J Mol Evol, 1984, 21 (2): 112-25
53. Schwartz RM, Dayhoff MO. Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science, 1978, 199 (4327): 395-403
54. McLachlan AD. Repeating sequences and gene duplication in proteins. J Mol Biol, 1972, 64 (2): 417-37
55. Fitch WM. An improved method of testing for evolutionary homology. J Mol Biol, 1966, 16 (1): 9-16
56. Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res, 1992, 20 (Suppl): 2019-22
57. Wilbur WJ. On the PAM matrix model of protein evolution. Mol Biol Evol, 1985, 2 (5): 434-47
58. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, 25 (17): 3389-402
59. Pearson WR, Lipman DJ. Improved Tools for Biological Sequence Analysis. PNAS, 1988, 85: 2444-8
60. Pearson WR. Rapid and Sensitive Sequence Comparison with FASTP and FASTA. Methods Enzymol, 1990, 183: 63-98

第四章　基因组分析

基因组研究的意义在于，它可以支持和推动生命科学中一系列重要的基础性研究。如基因组遗传语言的破译，基因的结构与功能关系，生命的起源和进化，细胞发育、分裂、分化的分子机理，疾病发生的机理等。在生命科学中，基因已成为共同的语言和基础。从整体水平研究基因的存在、结构、功能及其相互作用，从研究策略上把遗传学升华至基因学和基因组学，这在理论上具有深远的指导意义。在自然科学史上，人类基因组计划是第一次将人的物质结构、功能及其相互作用（关系）转换为信息的科学实践。它建立了遍布全球的不断扩充的数据库和信息网络。这不仅使生命科学开始了信息化革命，产生了极具生命力的生物信息学，而且也大大刺激了其他相关学科与技术领域的发展，如计算机科学、材料科学等新兴学科和数理化等经典学科，并将带动起一批新兴的高技术产业。其研究成果可直接指导和转化为实际应用，具有不可估量的社会效益和经济效益。本章介绍基因组分析的一些方法和生物信息学在基因组分析中的应用。

第一节　DNA 克隆和 PCR

DNA 序列的研究通常需要一些预测工具：包括对新发现的基因进行序列相似性分析以得到一些生理功能和结构的信息，发现有意义的序列片段（判断序列的生物学意义是通过预测生物学完成的）；检测基因或 mRNA 的分布情况（其分布情况通常是生物体中基因活性的指示剂）；PCR（聚合酶链式反应)用来扩增一定数量 DNA，以便用于纯化、测序和突变分析。

生物信息学基本上是对数据库的利用，即摘录、分类和分析基因、基因组和蛋白质的序列信息。这些信息的来源实际上就是基因测序。基因首先要被分离出来，克隆到适当的载体中以便于在实验室中操作；克隆和测序本身并不是生物信息学的内容，而是与生物信息学相关联的操作过程。序列数量的增加促进了统计分析的质量，同样，新的生物信息学软件的发展可以促进与序列相关的生物学功能的识别，也可以加快新基因的检测和克隆。

由于分析序列和结构信息所要求的准确率很高，所以要严格及时地评估所采用方法的准确性。这也有助于人们理解获得 DNA 和蛋白质序列的生物学背景。

一、 DNA 克隆

克隆一般是指单个祖先通过无性繁殖得到遗传上完全相同的一组细胞。本书所谓的克隆特指通过 DNA 重组技术操作 DNA 的过程，用来产生多拷贝的单基因或 DNA 片段，将所需基因或 DNA 片段从染色体位点上切割下来，并插入到可以在宿主生物中复制或扩增的载体 DNA 中。这些克隆载体源自病毒、细菌或酵母的 DNA 分子，包含着可以使 DNA 在宿主细胞中独立复制的启动子序列。细菌启动子用于哺乳动物表达系统的载体时，细菌 RNA 多聚酶会特异地控制载体 DNA 而不会影响细胞基因组。外源 DNA 片段可以插入载体 DNA 而不会使载体失去在细胞自然环境下自身复制的能力，从而使外源 DNA 片段可以在宿主细胞中大量地复制。载体可以是质粒（来自细菌）、粘粒（来自病毒）、酵母或细菌人工染色体（YACs，真核细胞来源；BACs，~150kbp 插入子）。具有基因表达调节元件的载体也称为表达载体。这些元件可以用来合成大量的 mRNA 或蛋白质，而宿主生物体在正常时可以是不包含或不表达这些基因的。

每个序列在储存于计算机数据库之前，首先要克隆 DNA 和构建组织标本的文库。在基因组计划中，基因组 DNA 通过鸟枪法的技术被机械地剪切或被放射诱导打断，将 DNA 片段收集起来

并重组到载体中，这些克隆片段的集合就组成了文库。一个载体包含一个基因，可以进行功能研究或转化细胞。

二、转录谱

如何选择被研究的基因？一个基因是一段碱基序列，抛开基因序列可以用于预测蛋白的功能不谈，基因序列对研究基因编码蛋白的生物化学和生理学的工作是必须的。在着手细胞生物学和生物化学的实验室研究之前，一个基因的活性谱（即基因在机体内的何种发育阶段和在何种细胞中表达并合成蛋白质）是首要的信息，以便于一个研究计划的设计或确定要研究的靶基因和药物。在特殊情况下，有些基因可能会限制在某些细胞类型、组织或器官中表达，它们的表达活性可能在健康和疾病（如肿瘤）的情况下有变化，或在年轻人和老人之间有所不同。

要评估一个基因在特殊细胞类型、组织或器官中的功能意义，首先要找到能表达靶基因的细胞（可通过 Northern blot 方法）。这个工作就是要找到作为蛋白质合成模板的信使 RNA（mRNA 中的 m 是指信使的意思，是指用于指导 DNA 序列翻译成为氨基酸序列的 RNA 的序列）。细胞内胞浆 mRNA 水平是很好的基因活性指示剂。mRNA 的高水平表达通常表示蛋白质表达水平高，但由于还存在转录后调控的过程，因此情况并不都是这样。在研究中，如果要涉及蛋白质的表达水平，必须另外独立检测。

确认 mRNA 表达的方法是利用放射性标记的寡核苷酸或 cDNA 探针进行杂交，寡核苷酸或 cDNA 探针以序列特异性的方式来识别目的 mRNA 。很明显，首先要获得一些序列信息，这些信息可以来自蛋白质片段或肽段的短氨基酸序列，或者通过查询 DNA 数据库得到所需要的序列。例如：一个与已知小鼠或大鼠基因同源的人类基因、或序列相似但并非同源序列且可能代表一个新基因的序列等等。比较细胞或生物体的不同标本在生命周期中的不同阶段表达谱的不同，或者比较在各种情况下细胞和机体发育之中细胞分化前后表达谱的变化，其比较结果可以构建一个时

间-空间图来表示机体内一个或一组特异基因表达活性的情况。

一旦确认了感兴趣的研究基因，就应当分离并扩增包含该基因的 DNA 片段。一种策略是利用逆转录酶制备 mRNA 片段的一个 DNA 拷贝。编码 mRNA 的基因可以在体外合成，称为互补 DNA 或 cDNA（complement DNA）。cDNA 代表基因的编码序列，包括mRNA 两端短的非编码区中的调节序列。明确真核生物基因编码区非常重要，这是因为大部分真核生物基因在细胞核染色体上的组织方式很特殊，与 mRNA 上的序列有明显不同。一个真核生物基因的基因组序列通常比其 cDNA 序列长，这是由于基因组序列是由编码区（外显子）和非编码区（内含子）组织而成的。尽管全部基因序列（包括外显子、内含子和调控序列）都被转录为 mRNA，mRNA 仍然会被催化修饰并去除内含子，这就剩下一个变短的 mRNA——包含了基因组序列中的所有外显子。这就是为什么要用 mRNA 来合成 cDNA 并且用 cDNA 来代表基因（与基因组序列明显不同）的原因，cDNA 可以被克隆到载体中并很容易地用于实验室研究（例如：体外合成蛋白质，用 DNA 转染细胞系和转基因动物研究）。

三、定点克隆

检测遗传性疾病基因的一种策略是定点克隆。克隆一种能导致某疾病或在疾病发展过程中起作用的基因，首先要利用遗传标记定位于染色体。遗传标记是基因组中很容易检测的短的非编码区。在这种方法中，要分析人群遗传学中的家族史——在人群中一些基因（等位基因）的突变以特异的频率出现。等位基因是指存在于某一人群中每个个体基因组中的一个特殊基因。基因的实际序列在每个个体之间有可能由于随机发生的突变而不同，而许多突变对于表型（也就是蛋白质功能）没有明显的影响，但有的突变可以导致蛋白质功能的改变，所以某一基因可具有若干种不同的形式，这种同一基因的各种不同形式互称为等位基因。一旦染色体定位完成，带有大的插入子的克隆可以通过物理作图来确

定，然后通过测序来确定基因。最后，通过突变分析比较来确认人群中受影响和未受影响个体的基因变化。

如果全基因组序列已经知晓，就可以大大地加快确认疾病相关基因突变的过程，这就是人类基因组计划要达到的目的之一。人类基因组的序列实际上只是一些个体的序列，只代表了某一人群中等位基因变化的少部分信息。研究等位基因的变化要通过比较部分表型的部分基因组来达到目的。由于不可能将每个个体的全部基因组测序，在定点克隆方法中对人群中不同个体的突变分析仍然是必需的步骤。多态性数据库就是为了产生这些信息而构建的。另外，有关各种疾病、感染的易感性、肿瘤和可能的机体生理代谢途径等等内容的数据库，在将来会越来越多。

NCBI 提供了 OMIM 数据库（Online Mendelian Inheritance in Man ，人类孟德尔遗传在线数据库）。该数据库是一个人类基因和遗传性疾病的目录，由 Victor A. McKusick 博士和 Johns Hopkins 大学的同事编辑，包含了文本信息、图片和文献信息。在 1998 年 9 月升级的数据库中有一个与眼睛散光相关的基因信息（OMIM 编号# 603047）。这个研究说明确定一个家族性疾病是非常困难的。散光是利用 OMIM 数据库进行研究的一个例子。

Clementi 等人研究了一个地区性的样本（Clementi，M. et al. Inheritance of astigmatism：evidence for a major autosomal dominant locus. Am J Hum Genet. 63：825-830，1998），其中 125 个家族的个体有眼睛散光，并有遗传史。他们利用 POINTER 和 COMDS 软件进行了复杂的分离分析。POINTER 不能区分不同的遗传模型，只能排除非家族遗传的假说。加入几个严格的参数后，COMDS 的分析结果确定了角膜散光的遗传模型，并且提供了该疾病是单个主要位点遗传的证据。这些结果提示遗传连锁分析是可行的，并且样本应当限制在具有严重受累个体的多个家庭。考虑为常染色体显性遗传疾病比较合适。

在应用新开发的软件进行这个基因组分析之前，散光被认为

是没有遗传性的，而环境因素是其主要的致病因子。例如，在 1989 年之前，Teikari和 O'Donnell（Teikari，J. M. O'Donnell，J. J. Astigmatism in 72 twin pairs. Cornea 8：263-266，1989)就提出遗传因素不是散光的致病因子，而环境因素是其主要的原因。

四、多聚酶链式反应（PCR）

DNA 扩增的革命性技术革新是多聚酶链式反应（Polymerase Chain Reaction，PCR）。1985 年由 Kary B. Mullis 开发了 PCR 技术，他后来在 Cetus Corperation 就职。1993 年，由于对分子生物学的突出贡献，Kary B. Mullis 获得了诺贝尔奖。今天，几乎全球的每个分子生物学实验室都在使用着他的技术。其操作过程已经实现自动化，扩增从小量到大量 DNA 的自动化仪器都已经商品化。整个过程从引物设计（设计寡核苷酸以寻找基因组文库中的靶基因序列）到研究器官中的基因表达都可以用计算机软件完成。

由于序列数量的增长使我们可以寻找未知基因的功能单位。通过测序信使 RNA 来大规模地确定基因的表达，使研究者可以跟上基因组计划测序结果（包括公共的和私人公司的文库和数据库）的步伐。DNA 序列可以用来产生一些短序列，这些短序列又可以用来检测 mRNA 。为了提高寻找较好的药物靶分子的效率，制药公司开发了微阵列技术和 DNA 芯片技术，这种技术可以在一个实验中扫描成百上千种基因片段。DNA 芯片技术由加利福尼亚州 Santa Clara 的 Affymetrix 公司（http：//www. affymetrix . com/）开发，DNA 芯片技术是检测组织中基因表达分布和表达序列标签表达分布的领先技术。

PCR 对于生物信息学，尤其是对于基因组计划的重要性，在于它可以在序列尚无任何生物学信息时使 DNA 得以扩增。这意味着只要已知一段短序列（10-20 个碱基长度）就可以扩增编码区或非编码区。由于这种技术是在非细胞环境下利用酶控制 DNA 的扩增，所以对于极少量的标本也非常敏感。

五、发展中的测序技术

人类基因组计划的一个主要焦点就是自动测序技术的发展。自动测序技术可以在一天内准确地测定出 10万个碱基,检测每个碱基的成本平均少于 50 美分。包括测序和检测技术发展在内的一些特别的目标，要求这项技术更快、更敏感、更准确和更经济。现在，许多测序新技术已被开发出来，最有前景的一种将会被开发并广泛地应用。第二代的测序技术将会使测序的速度和准确率提高 10倍，同时降低检测的费用。一些重要的疾病基因可以用高电压毛细管技术或超薄电泳技术来测序，以增加片段的分离率。应用共振离子质谱的方法可以检测稳定的同位素标记。第三代无胶测序技术的目标是提高几个数量级的测序效率，并应用于大部分人类基因组测序的工作中。这些技术包括增强每个标记碱基的荧光检测，利用扫描仪或原子显微镜来直接阅读 DNA 条带中的碱基序列。DNA 序列的增强质谱分析，可利用已知基因的短片段来杂交测序。大规模的测序计划将为促进当前技术的进步提供机会，也使互相竞争的研究人员面临更大的挑战。

荧光标记 DNA 片段大大地加快了应用 Sanger 双脱氧链末端终止法测序的速度。这种方法是利用了酶能够促进 DNA 合成的能力。通过加入可以终止延长过程的核苷酸底物，产生不同长度的 DNA 片段，细致地分离这些长度只有一个核苷酸差别的 DNA 片段，我们就可以读出克隆 DNA 的全部序列。

另外，Maxam-Gilbert 方法是利用酶去降解 DNA 克隆的特殊位置的碱基，从而生成不同长度片段的混合物。然后利用凝胶电泳来分离这些长度只有一个碱基差别的片段。

六、监测测序进展

许多网络站点包含了各种基因组计划和数据库的信息，并且有超级链接，它们都包含了相对完整的特殊软件。但所应用的数据库仍需要科学家来验证其数据库的数量和及时性。能较好地监测人 DNA 克隆测序进展的例证是 Sanger 测序中心的站点（www.sanger.ac.uk/HGP/stats.shtml）。该站点也提供了 FTP

站点的超级链接，可以显示一个克隆或序列的 FASTA 序列格式，同时还包含了该序列的摘要（可以确定该序列是否就是要研究的蛋白或基因，是否与其他物种同源）。

监测测序进程是一个有趣的活动。大量的序列数据极快地加入到数据库中的速度是惊人的。“进程统计(Progress Statistics)”(http://www.sanger.ac.uk/Info/Statistics/ ）可以在网上得到，其中显示了英国医学研究理事会（British Medical Research Council）Sanger 测序中心完成的和未完成的核苷酸序列。未完成的克隆提供了不完整序列的更新信息。这使得我们可以快速地得到感兴趣的新基因。这些序列信息在应用时必须小心，因为其中可能有一定的错误，应当将其看作是未发表的序列。在这里要注意的是，这些克隆信息只是指 Sanger 中心的克隆序列，并不代表任何其他物种的全部序列数量。完成的克隆可以经注释后提交到 GenBank、EMBL 和 DDBJ（日本 DNA 数据库),未完成的数据则不行。

一般地，由于人们只研究自己感兴趣的课题。所以，对互联网上的信息，不同的人取舍不同。其他各个组织建立的站点与三个主要的公共数据库 NCBI、EBI 和 NIGJ（National Institutes of Genetics in Japan，http://www.nig.ac.jp/home.html）不同，他们只概括地反映了在他们工作范围以内的信息。有些人集中研究特殊物种的单个染色体；有些则对全部染色体的作图和资源开发感兴趣；有些人集中在自动化数据处理和分析；还有些人对开发新软件感兴趣，如用于分析序列、比较基因组、研究基因的结构和表达、确认多态性和研究与功能相关的染色质结构等方面的软件。所有这些研究加起来将加深我们对基因组生物学功能的理解。

第二节　DNA 序列分析的计算机工具

在生命科学中，计算机发挥作用的经典例证是测序、序列分

析比较、追溯进化和突变、为药物设计发现序列相似性、预测蛋白质功能、预测基因在细胞机制和疾病发生中的作用。

集中式数据库（centralized database 的用途不仅使科学家们知道彼此克隆研究的结果，而且可以作为他们比较遗传学研究的基础。没有不同物种间不同蛋白质的 DNA 序列关系信息，就不可能理解进化。生物信息学的工作，总的来说，是处理序列数据库中的文献和生物学注释、支持利用软件进行序列排列、确认基因、将 DNA 序列翻译成氨基酸序列、查找同源序列（进化相关序列）。这也就是收集、储存、组织和注释原始序列，构建二级和三级数据库。

十五年前，研究人员通过电话给对方读 DNA 或氨基酸序列是很常见的事情。这样就导致了一个人为的“突变率”，估计这个“突变率”远远超过了在自然状态下 DNA 复制和转录过程中基因的突变率。而今天，从 GenBank 和 SWISS-PROT 数据库中下载一个文件非常简便快捷，而且几乎可以避免此类错误的发生。

一、数据库数据提交

集中式数据库中序列信息的主要来源是科学家本身。现在 Internet 的发展使提交信息到 NCBI、EBI 和 DDBJ 的过程非常简单。BankIt（在线序列数据提交工具）或 Sequin（个人使用软件）由 NCBI 提供，用于向 GenBank 的工作人员提交序列信息和生物学注释。由 GenBank 的科学家给予每个信息附加序列号（accession numbers），并很快公布在公共数据库中（通常在 48 小时内）。GenBank、EBI 和 DDBJ 每天互相交换新提交的数据以保证每位科学工作者投送的信息是非冗余性的（指序列只投送一次）。

序列作者可以更新他们的原始信息。在正常情况下，科学家们各自发现的基因只有一个序列和一些相关生物学信息。而对基因组计划来说，来源于 ESTs（expressed sequence tags，表达序列标签）、STSs（Sequence tagged sites，序列标签位点）和 GSSs

(genome survey sequences，基因组检测序列）的序列信息的提交需要特殊过程，这些序列与传统的功能基因和蛋白序列不同，他们相对较短而数量巨大。

ESTs 是长为 300-500bp 的短序列，它们代表表达基因，通过提取组织或细胞的 mRNA 然后经过反转录获得。除它们正确的序列外，这些短的序列标签对于在染色体上定位基因很有帮助。EST 序列的提交通常包括序列和作图信息，通常这些信息一次以十个到上千个的批量提交，在引证信息、提交数据和文库信息上有冗余性。GenBank 在线提供了提交序列所需要的信息。

STSs 与 ESTs 在长度上相似，并且每次提交的数量也相似。它们不代表基因表达的情况，但是用 PCR 确认基因组时它们是单一的标志物。尽管 ESTs 是公共数据库中数量增长最快的一个子数据库，但由于基因组中大部分序列为非编码区， STSs 将会在数量上超过 ESTs。

由于基因组序列对科学机构有潜在的用途， NCBI 每天都处理提交的基因组序列信息。其中包括基因组中心、克隆名称、序列号等，基因组序列也可以在未完成之前直接提交。 NCBI 将高通量基因组（High Throughput Genomic，HTG）序列分为三期：1）未完成，未排序；2）未完成，已排序；3）高质量的已完成的序列，不包含任何序列间隙。由于高通量测序和提交步伐的不断加快，确认其中的错误是很重要的环节。

为了促进这个过程，NCBI 建立了流水线式提交程序和序列公布的最后期限，以保证能够快速和无错误地将新序列公布在它的 ENTREZ 系统。没有一定的速度和准确性，任何数据分析都很困难。由于许多生物信息用于预测（确认新基因、新功能、药物设计、预测结构和进化树关系分析），错误会很快地随电子媒介传播，序列水平的错误会导致解释和结论的错误。 NCBI 最关心的是对储存信息的错误注解。为了解决这个问题，受过专门训练的科学家必须整理数据库并纠正其中的任何错误。另外，错误注解的

传播也降低了比较生物学数据的可靠性。

尽管 GenBank 的序列主要依赖各位生命科学家和高通量测序中心（如 Sanger 测序中心、TIGR 等）的直接提交，但 NCBI 的工作人员还需在生物医学杂志中查找发表的序列和结构信息用于对序列的注解，正如 GenBank 96.0 公布版本所写：

> GenBank 包含由作者直接提交的序列，也包括 NLM（National Library of Medicine，美国国立医学图书馆，http://www.nlm.nih.gov/）通过浏览生物医学文献制作的部分材料。NLM 每年要从 3400 种杂志的 325,000 多篇文章中查找序列数据，这些数据列在植物学和兽医学杂志的附录中，这些杂志与国立农业图书馆（National Agricultural Library）有合作。GenBank 是美国、欧洲和日本的三个国际协作数据库的组成部分；欧洲的协作数据库是欧洲分子生物学实验室（European Molecular Biology Laboratory，EMBL），位于英国 Hinxton Hall；另外还有日本 DNA 数据库（DNA Database of Japan，DDBJ），位于日本的 Mishima。协作数据库中的序列数据也与基因组序列数据库（Genome Sequence Database，GSDB）有合作，基因组序列数据库位于新墨西哥州的 Santa Fe。专利序列由美国专利和商标局安排与三个协作数据库合作，并与其他国际专利局通过国际数据库合作。数据库转换为各种输出形式，包括普通文本和 ASN.1 版本，ASN.1 形式的数据包括在 Entrez 中，序列的 CD-ROM 也可以得到。通过匿名 FTP 可以得到普通文本（ftp://ncbi.nlm.nih.gov/genbank/release.notes/gb96.release.notes）.

为了理解今天庞大的计算机网络和遗传学数据的巨大流量，我们必须回顾分子生物学处在婴儿期的 40 年前。那时，许多今天拥有的技术还没有出现，遗传密码也刚刚发现，随后不久发现了限制性内切酶（切割 DNA 的工具）。在当时，蛋白质的测序速度比核酸的测序速度快。但生物化学家也需花几个月到几年时间才能通过顺序降解大量的纯化蛋白质，来搞清楚一个蛋白质的氨基酸序列。生物化学的先驱如 Margaret Dayhoff，她是最早将氨基酸序列比较用于进化分析的生物学家之一。她首先认识到建立公共

序列数据库的必要性，她的观点对于发展基于计算机基础上的分析工具是十分重要的。在 Margaret Dayhoff 的努力下，60年代早期建立了第一个蛋白质序列数据库。

今天，氨基酸序列可以按常规通过分子生物学的方法得到。也就是说，首先基因测序，然后应用适当的密码子规则从 DNA 序列推断出氨基酸的序列。

然而，氨基酸测序目前仍用于分析短肽。这一方法的作用最近被越来越热门的蛋白质组学夸大了。从蛋白质表达谱获得的肽段经过微测序确定其分子量和电荷（等电点），在这些蛋白质提取物中获得的短氨基酸序列的基础上，可以很快地分析蛋白质表达谱和翻译后的修饰情况。

直到 1980 年，DNA 序列数据库才开始建立。由于科学团体之间需要快速和可靠的信息传递，国际互联网应运而生。在应用网络浏览器之前，从远端计算机下载文件的标准方法是通过文件传输协议—FTP 和 Kermit，即使用公共软件包。这些方法目前仍然应用于超级计算机之间的通讯和文件的上传下载（见 Pittsburgh 超级计算机中心；http://www.psc.edu）。在 1989 年以前，最通用的序列提交和查询形式是普通邮件（硬盘拷贝、软盘和磁带）、电传以及拨号在线网络。人类基因组作图文库（HGML，冷泉港实验室）用手工更新它们的数据库注解， GenBank 的科学家们则手工扫描各种杂志以获得已发表的序列。在那一时期，只有 50% 的记录是由相关领域的科学家直接提交的，而其中的 70%是以计算机可读的形式投递到以 UNIX 为基础的 SUN 工作站。当然，UNIX 至今还没有被取代。网络浏览器是适用于 Windows 和 Apple 操作系统的计算机程序，具有友好的界面，把 PC 机转换为类似在 UNIX 操作系统运行下的超级计算机和工作站的终端。

虽然超级计算机有花费高和访问限制等缺点，但有时在局部站点中使用许多独立的软件程序或可以下载的数据文件时有很大的优势。许多生物技术公司在局部服务器上建立镜象站点将公共

数据库和专有数据库结合起来，以利于数据储存。这样做的结果，就将他们正在进行的生物信息学研究用于商业运作。对于个人用户，通过网络浏览器在远端计算机上进行交互式分析，并应用 e-mail 接收结果是生物信息学应用最普通的形式。个人计算机的计算能力和速度的提高使独立程序的应用更为可行，并且减少了对大型工作站和远端服务器的依赖。

在 1983 年到 1988 年这 5 年期间，从 DNA 序列的发表提交到数据库中可以获得的平均周期从 1 年下降到 5 个月。1988 年是人类基因组计划启动的标志年，那时在 GenBank、EMBI 和日本的序列数据库中包含了 1200 多种生物的序列数据。而今天网络站点的记录和查询形式可以提供即时的服务。十年以前，订阅 GenBank 数据的用户，每 3 个月可以收到记录数据的磁带；当时 CD-ROM 的技术刚刚开始应用，但是花费昂贵，限制了其广泛的应用。EMBL 数据库的一年磁带订阅费用是 200 美元，而美国非商业用户的 CD-ROM 订阅费用是 400 美元（Methods Enzymology，1990，vol. 183，p. 29）.

二、数据查询

序列分析包括 4 个主要的生物学相关主题：1）比较基因序列以得到相似性信息和从进化树分析中确认同源性。2）确认基因的基因组结构，包括开放读框、外显子-内含子分布和调节序列。3）预测蛋白质的结构。4）基因组作图，染色体上基因的线性排列和在代谢途径中的作用评估。

当前可以获得的 DNA 和蛋白质序列的数量非常大，查询信息可以形容为“挖掘生物学数据矿藏”。搜索引擎执行的两个基本任务是：对储存信息检索的简单字符串搜索（如： GenBank 的核酸蛋白查询；PubMed 的 MEDLINE、三维结构、基因组和分类数据库查询等），检索、排列和比较序列或结构的相似性查询（如 BLAST）.

序列分析的第一步包括以一定的标准来检索序列（其中之一是查找序列的相似性和一致性），这可以通过诸如 BLAST 的搜索

工具完成。如果没有已知序列，NCBI 的搜索引擎可以在核酸或蛋白水平通过相应的蛋白名称、研究目的蛋白的作者姓名或序列号扫描数据库。这些搜索可以在选择的数据库中检索到数据文件中包含检索词的相应记录，包括记录的编号。

例如，如果一个研究人员拟寻找第四军医大学病理学教研室发现的名为“dif14”的基因的核苷酸序列，在 Entrez 查询站点的“search”选项中选择“GenBank”（http://www.ncbi.nlm.nih.gov/Entrez/），他可以简单地输入关键词“dif14”，点击“Go”，就可以找到相关的记录。

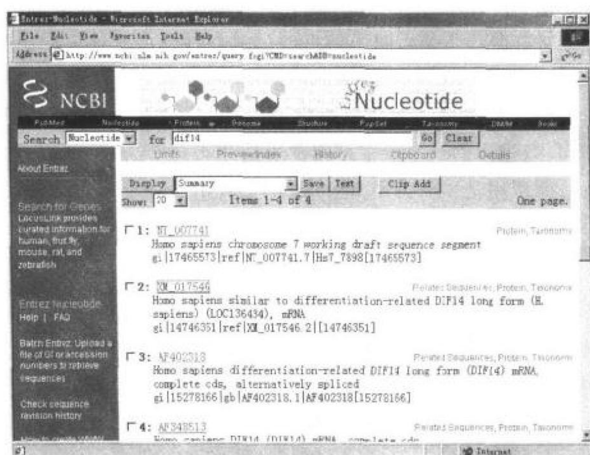


图 4-1 Entrez 的核酸查询，查找“dif14”的结果（2001 年 12 月）

查询结果出现的窗口显示有 4 个相关文件（图 4-1），可以通过点击条目链接来显示每个查询结果。感兴趣的研究者可以得到其序列信息（图 4-2，FASTA 格式）、注解信息(GenBank 报告)、图形显示（Applet Java 图形）以及相关蛋白或核酸序列，如果该序列有相关文献，还会显示文献链接（MEDLINE）。图 4-1 中可以看到最后一个记录的 accession number 是 AF348513，指 dif14 基因的短片段形式，这可能是该基因 mRNA 在不同剪切方式下的

图 4-2　序列号为 AF402318 的 dif14 查询结果（FASTA 格式）

产物之一。而 AF402318 条目是 dif14 的长片段形式。该基因预测的开放读框编码一个有 490 个氨基酸的蛋白质，经蛋白质跨膜区预测软件分析，它可能是人类的一个九次跨膜蛋白质。选择显示"FASTA"链接可以显示该序列的 FASTA 格式。查询结果中包含序列号为 NT 007741 的条目，它是 7 号染色体的工作框架图（working draft）序列之一。由于 dif14 的序列定位于 7 号染色体，并且位于 NT 007741 条目序列之中，所以查询结果中也包含了该条目。

选择"protein"链接可以显示人 dif14 长片段蛋白质的记录（图 4-3，accession number AAK94061），FASTA 链接显示该蛋白质的氨基酸序列。

氨基酸序列的 FASTA 格式可以用于 BLAST（BLASTp search）同源查询非冗余(non-redundant）数据库(GenBank CDS translation ＋ PDB ＋ SWISS-PROT ＋ Spupdate ＋ PIR ＋ PRF，共 855，480 个序列，269，408，271 个字母），以查找相关序列 这种 BLAST 查询可以通过在 FASTA 格式窗口选择并拷

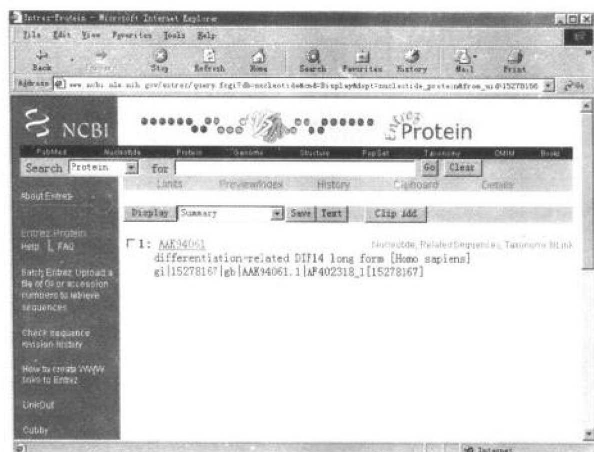


图 4-3　Entrez 蛋白质查询 dif14 的结果（AAK94061 是序列号为 AF402318 核酸预测的编码蛋白）

贝氨基酸序列，粘贴到 Standard protein-protein BLAST [blastp] 查询窗口，选择 blastp （由于这是一个蛋白质序列）来完成。

查询结果有 26 个，其中包括小鼠肢体发育相关蛋白质 LMBR1、人 lipocalin-1 的膜受体蛋白质、鸡 Saph1 蛋白质、线虫的 R05D3. 2. p 蛋白质、人脑 my034 蛋白质和一些未知生物学意义或未命名的蛋白质（2002 年 1 月 24 日查询结果）。

得到的蛋白质相似性水平也提示了物种和分类间可能的生物学关系。序列的同源性高低由 E 值来衡量,E 值表示随机命中的几率,如果 E 值是 0 或接近 0(如 BLAST 查询结果中 AF402318 为 0)表明是不可能随机碰到的，也就是说同源性非常高。同源性序列通常是在一定有意义的界值下显示的，通常来说如果查询同源序列将 E 值设定为 0.1 是比较合理的,E 值大于设定界值的序列表明与查询序列没有关系。但即使两个基因的 DNA 序列没有多大的同源性，他们的氨基酸序列和蛋白质结构水平也有同源的可能。随着得到的高解析度蛋白质结构数量的增加，可以肯定的是

· 数据库中的相关序列
· 结构预测／与 x 线衍射结构的比较
· 功能未知时，开放读框 ORF
· 结构域部分
• 跨膜部分
· 信号肽序列
· 糖基化位点、磷酸化位点以及在脂质中的锚定位点
· 选择性的术语命名
· 诸如调节序列的遗传学信息
· 翻译
· 2 维凝胶电泳、等电点（电荷）、分子量
· 参考文献

启动基因组计划的原动力，来自于人们对代表着或包含着基因的 DNA 序列的确认需求。基因是各种生命体基因组中的功能单位，它包括调节序列和位于启始密码子和终止密码子之间的开放读框。开放读框决定其相应蛋白质的氨基酸序列。不同生物之间的基因结构有显著的差异，而且存在着两大类型：即拥有连续开放读框和拥有间断开放读框这两种不同的结构（外显子和内含子；所有外显子共同代表开放读框，而内含子则在 mRNA 水平被酶切掉——即 RNA 剪切）。后者只存在于高级的生物体内（真核生物），在细菌与原生质中则没有。

五、开放读框和未确认读框

如果一个基因已经测序，但没有相应蛋白质的信息，就不会有相应的生物学功能的信息。DNA 序列是在基因组计划中获得的最原始的结果。这样，就必须对 DNA 长长的重叠序列进行分析，以期找到存在着的基因。完成这一工作，需要借助软件来确认起始密码子和终止密码子之间的开放读框（Open Reading Frames，ORFs）或未确认的开放读框（Unidentified Reading Frames，URFs）.

ORF 的长度与编码蛋白的大小、分子量密切相关，并且其长度是确认 ORF 的有用指征。在真核生物基因中，剪切位点（间隔着外显子和内含子的位点）有鲜明的特征，它在确认基因时提供了辅助的作用。由于基因是一个功能单位，在临近起始密码子的位点存在着共同序列。

许多网站提供了可以分析一个 DNA 序列中 ORF 存在与否的各种工具，它们允许对蛋白质的相关氨基酸序列及可能的结构特征进行预测。如果经过序列排列找到了相关序列，而这个相关序列中包含了一个基因的序列，这个结果是一个基因可能有生物学功能的很好的标志。

ORF finder 是一个有效的图形界面分析工具，它可以在用户提供的序列或数据库的序列中找到大小不同的所有开放读框。这个工具是通过使用标准密码子或其它一些特殊物种的密码子，来确认所有的开放读框的，推断出的氨基酸序列能以多种形式保存；而且，它可通过 BLAST 服务器来查询序列数据库。ORF finder 对于提交完整而精确的序列是相当有用的，这里还包括了 Sequin 这个序列提交软件（摘自 http://www.ncbi.nlm.nih.gov/gorf/gorf.html）

为了保证能正确地预测新基因，必须仔细地选择使用物种的密码子。NCBI 也提供了一个密码子使用数据库（http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html）。这个数据库包括所有真核生物的标准密码子及其分类分支。

ORF finder 搜寻 cDNA 序列，以寻找在起始密码子和终止密码子之间适当的伸展序列。所谓“适当”是指一个基因的大小及蛋白质的大小是适当的（专指功能未知的蛋白质），或可以通过同源序列推断出来。在后者的情况下，ORF finder 可以预测新基因编码蛋白的功能。它提供了确定被研究的 cDNA 序列是否包含基因的功能位点的方法。ORF finder 对于筛选细菌基因组、cDNA 文库和 EST 数据库是非常有用的，但它不能分析真核生物的原始

序列。因而，目的基因的片段，即外显子，首先应被分离出来，然后克隆、测序，放在一起组成重叠序列。重叠序列可以包括基因的连续编码序列。

基因是染色体上的工作单位，包括 ORF 片段以及对基因表达的调节非常重要的非编码区。真核生物基因结构常常是很复杂的，含有重组过程。在重组过程中，一个基因可以各种复杂的方式重组外显子而导致生成不同的基因产物。这就是免疫球蛋白的高可变区结构域形成的基础。在 DNA 片段中确认基因的软件可以从 Bylor 大学医学部获得（Gene Finder，http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html）.

第三节　基因组分析

基因组分析可以确定基因在染色体中的位置，并提供相关信息。这包括：与其他基因的遗传、连锁，在遗传学、医学方面的重要性，基因治疗，示踪常染色体突变及 X 染色体连锁疾病等。

例如：酵母蛋白质组学数据库（YPD）将 DNA 序列、蛋白质结构和功能、细胞内定位和通路以及细胞周期信息与一个连贯的数据库相连。这个数据库又与相关的文献信息相连，而且具有将数据获得权售与公司的商业目的。可用于蛋白质组学与基因组学的比较、二维凝胶电泳分析、图象处理、储存查询以及特征查询等（Virage Inc. www.virage.com）.

一、基因组的组织

生物信息学工具和数据库慢慢地融为一个反映有机体复杂性的整和系统。随着一个个小生物基因组计划的完成，对三种生物基因组（原核生物、原生质、真核生物）的不同点的理解使得一个有机体的组成及功能与基因组组织的关系也渐渐清晰了。真核与原核生物在基因组结构上很不相同。编码区与非编码区有着不



图 4-4　人类基因组资源网页

同的出现频率。细菌基因组排列紧密并有很少的非编码区 DNA、真核生物染色体通常非常巨大，而且有大量的非编码 DNA，尤其是植物。真核生物和原核原生质的基因常分裂为不连续的片段，称为“外显子”。

包含生物全基因组的特殊数据库提供基因组中基因相互关系、空间位置相邻信息和共调节等信息。例如，在一些菌种中有额外的酶催化步骤，而在其他菌种中没有。发现这一现象的方法就是观察是否这些特异蛋白属于一个基因簇（这个结构为操纵子），并沿着微生物的基因组排列在一起，就象氨基酸合成的整个路径的酶排列接近，并且以协同方式调节，以免每个酶的单独控制都需要一个通路。了解通路的存在及编码通路中每一种酶的基因对于理解突变如何影响细胞生理有重要意义，通路中一个环节的酶的突变影响到整个途径，因为它组成一个表型。原核和真核生物全基因组测序过程无疑有助于确定机体基因组结构在代谢过程中的生理功能及其重要性。

尽管基因非常重要，因为它们编码了细胞内全部蛋白质和

RNA，但这些结构基因只是组成基因组的一部分，尤其在真核生物中（如：真菌、植物和动物）。例如，估计 90%以上的人类基因组为非编码区。不久前，非编码基因还被认作是 DNA 垃圾，这反映了我们对其功能的理解和认识的不足。经研究越来越多的非编码 DNA 对蛋白质或 RNA（调节性、结构性和酶）合成起作用，并且认为对细胞 10% 的编码 DNA 有很重要的意义。这些非编码 DNA 在细胞特异基因的复制和表达方面是很重要的。它似乎包括“只读”信息，即与基因表达和复制相关的蛋白质结合位点的短序列。这类蛋白质是生长因子和激素受体，这些蛋白质结合部位对细胞来说是非常重要的，在细胞的分化、形态发育及胚胎发育时期起了很重要的作用。

在研究进化中的 DNA 时，从 DNA 的非编码区得到的启示是很巨大的。因为突变是随机事件，染色体的非编码部分包含了大部分碱基位置的变化，并且为染色体的重组和这种隐性突变的积累提供了“发挥的场所”。

多态性标记（用于 DNA 指纹技术中的标记)，在这部分 DNA 中可以应用。对于不同个体基因簇的遗传学研究结果反映了不同个体之间核苷酸序列的高频突变。它反映了被 DNA 限制性内切酶切开的 DNA 片段上序列的变化（限制性片段长度多态性）。这种遗传学多态性最近已被用于法医学的实践中。这种称为遗传学“指纹”产生的信息对于某个个体来说，是几十亿人群中独一无二的。PCR 成功地应用于身份确认，在犯罪现场中发现的极少量血样、坏死皮肤或一根头发都足够用于 DNA 扩增分析。

为了理解生命的“蓝图”和生命本身间的关系，我们需要了解基因在基因组中的相对位置信息以及蛋白质序列和结构之间的关系。由于蛋白质不是孤立的实体，多种蛋白质之间的相互作用是细胞活性的基础，单个基因的选择压力可能与几个有相互作用的蛋白质的基因有关。这使得序列—结构和结构—功能之间的关系变得相当复杂，多个蛋白质的相互作用更加复杂。现在，有诸

如基因组学和蛋白质组学的新技术，这些技术可同时确定多种 RNA 或蛋白质表达水平，是研究细胞内复杂的分子间相互作用的起点。

我们如何衡量遗传特征的独立性和相互依赖性？这可以参考 Gregor Mendel 以及他的有关豌豆颜色和硬度的分离遗传研究实验。在分子水平，两个独立表型是由位于染色体上的基因来编码的，这些基因在染色体上的位置是有一定距离的。如果两个基因位于同一条染色体上，他们通常一起遗传，也就是说他们不会分离。但这并不是一定的，因为同一条染色体上的基因间距离是极重要的。若两个基因间距较大，其分离的可能性亦增高。

组蛋白（histone）的分子进化研究证明了基因组结构与染色体稳定的重要性——组蛋白负责组装和储存 DNA，形成染色体的高密度形式。在细胞分裂期间，染色质浓缩成熟知的双臂结构，即一对染色体。但在细胞正常休眠状态下，染色体常松散分布，并且可在 RNA 聚合酶的作用下转录成合成蛋白质的 RNA 和其他转录因子。这就是基因调节（包括转录和表达）的本质，这是一个在 DNA 链和结构蛋白（或组蛋白）、核酸合成蛋白（聚合酶）和 DNA 结合蛋白（或转录因子）之间的动态平衡。这个动态平衡控制聚合酶与 DNA 分子的结合。

由于基因可以编码蛋白质，而蛋白质控制着细胞生命中的每个过程。所以基因的转录对机体活性的重要性是显而易见的。尽管染色体结构的重要性目前还不是很清楚，但有迹象表明改变染色质的结构对细胞是致命的。对组蛋白氨基酸序列的分析为此提供了一个证据。在真核生物中的各物种组蛋白高度保守，它们是动植物、真菌关键的一个遗传特征，它们保守的序列也提示着所有现代真核生物在进化上都来自一个祖先细胞或物种。事实上，组蛋白已被用作分子计时器或分子尺，用来在亲缘性很远的物种间测量进化树距离，即两个不同物种的分离时间长短。一个生物（或是一个群体）的存活与它的表型有关，而且表型的遗传变异发生在

DNA 水平并在随机发生突变的碱基位置上积累下来。如果表型对应的是一些致死性结果，那么突变就被拒绝。生物体或者在发育成熟前死亡，或变得没有繁殖力，这样就失去了把突变基因组传给下一代的机会如果一个基因的核酸序列没有被拒绝，那么随着时间延续，在一个基因上的突变积累速度就是一个直接测量标准。它能反映出个体生存能力表现型的重要性，但对群体则不然。然而在一个人群中，等位基因多样性是特异基因突变易感性的一个指标。组蛋白基因在数亿年中极低的突变率，表明这些蛋白的结构对所有真核生物来说都是必不可少的。这意味着在细胞周期不同阶段的染色体装配与基因复制、转录同样对细胞的生存是极其重要的。

通过把基因的组成、组织和瞬时表达情况的信息进行分类和总结，基因组学将给出细胞功能进化的细节内容。因此， Internet 成为科学家的一个极其重要的工具并不令人惊奇。因为网络中有众多的数据库，其中含有数千种的基因组信息、物种分类以及以进化树——“生命之树”(Tree of Life)——的形式显示的进化关系。进化树是理解进化关系的可视性方法。

“生命之树”以图形形式描述了地球上来自同一祖先的多样化的生命形式。人们相信只有“一棵这样的树”（也就是单个始祖细胞），即生命并非有多个起源。经过推敲并证实的一个观念是，生命是在很偶然的机遇下起源于非生命物质的。 Arizona 大学的生命之树工程（http://phylogeny.arizona.edu/ tree/life.html ）提供了地球上各种生命的可视性进化树。这不是一个分子形式的进化树，而是一个经典的分类学进化树。这个工具对没有经过进化学、动物学、植物学及生态学正规训练的分子生物学家是非常有用的。这个工程包括了地球上生物的多样性、历史及其特征的各种信息。这是由 Arizona 大学的 David. R. Maddison 创立并协调的一个有多位作者的网站。

蛋白质常作为大的蛋白复合物的一部分而存在，而且只有在这些大复合物的所有成分都存在的情况下，才能研究这些蛋白质

的活性。它们不是独立的，因此他们的基因也不可能是独立的。可有些蛋白质复合物是由随机分布的基因编码生成的，各组成部分在染色体上没有任何连锁的关系。在人类基因组中某一组基因如此明显地缺乏组织性，是否有什么意义呢？红细胞中的血红蛋白——其功能是把氧从肺转运到靶器官（如肌肉或脑），它由两个不同的基因编码的四个紧密结合的蛋白亚基组成。这两种基因称为 α 和 β 血红蛋白基因。有功能的血红蛋白复合物包含着每个基因产物的两个拷贝，形成正确的复合物需要这两种基因一同表达。由四个 α 亚基或四个 β 亚基组成的血红蛋白是没有功能的，编码血红蛋白 α 亚基的基因，实际上包括了序列上稍有不同的一组基因，它们在胚胎发育的不同时期表达一系列不同的蛋白。这样，在整个发育时期的某一特定时刻，只有一个拷贝的 α 亚基基因簇可以表达。α 亚基基因簇位于 16 号染色体上，不同拷贝的位置非常接近；而 β 亚基基因簇位于 11 号染色体上。

解剖学和生理学的表型是多基因表型，也就是说由几种基因产物组成基因型。除了个体外观上显而易见的特征外，细胞的新陈代谢是研究多酶反应途径的最好平台。像糖、脂肪、氨基酸、脂质等的合成与分解代谢是一些复杂的相互联系的代谢途径的一部分。在不同生物的基因组中构成每个代谢途径的基因组织是不同的。有一个原则，在功能和结构上相互作用的蛋白与它们的基因在染色体上的位置之间，没有严格的相互联系。有时这些基因在基因表达后紧密地结合成一个功能单位，但它们的基因却散在地分布于整个基因组，功能基因组学可能会解决这个问题。

在特殊的 DNA 序列和生物染色体形态之间有一个确定的关系。以下独特的形态学特性已被证实：

- 端粒区（串联重复序列；与衰老有关）
- 着丝粒区（串联重复序列）
- 核仁组织区（核糖体 RNA 基因，与染色体对中间的形态有关）

由于基因功能和染色体结构间的关系密切，基因组的物理图谱对于理解生物体的独特性及它的发育规律（生命周期）是极为重要的。一个生物体的独特性不仅取决于其基因的组成，而且取决于染色体结构。哺乳动物的染色体存在着中间着丝粒和近端着丝粒两种形式。已经发现不同物种的个体（尽管在基因序列上关系很近）由于染色体结构在细胞融合和分裂中的不相容（染色体结构即依赖于组蛋白形成超级结构），它们在生殖方面也是不相容的。在此，我们可以见到一个相互作用的环。编码组蛋白的基因是由这些蛋白质相互之间以及与 DNA 之间的相互作用来调节的。由于这种相互作用在细胞分裂中非常重要，又可以决定细胞的生存能力，所以组蛋白核苷酸的突变影响其氨基酸的组成。这又可以影响到染色体的结构，继而影响组蛋白基因的复制和表达。

二、基因组作图

基因组数据库在研究那些功能还没有明确的新基因的工作中所起的作用越来越大。通过类推的方法考虑一个基因的位点和与染色体位点的关系，就有可能推断其功能，并且对设计将来的实验很有用。染色体定位与 DNA 序列相似，常常有变化（如：突变），并且在每一代之间都可能有变化。在真核生物中，染色体片段的重排（同源重组、相互杂交、减数分裂和有丝分裂）是个体之间遗传多样性的重要部分。遗传多样性以染色体重排为基础。正象如上提到的，尽管全部基因组的内容还是稳定的（全部基因都遗传了），但个体在遗传上仍是各有其特点的。重排可以影响和改变基因表达的顺序和程序。

许多这样的重排过程也可以导致疾病，这也是理解基因表达和染色体形态之间关系的另外一个原因。与医学因素相关的基因组数据库内容逐渐增加的现象，也反映了这个研究目的。同样，有关遗传性疾病信息的站点数量也在逐渐增多（如 NIH 的健康信息：http://www.nih.gov/health/）。

1．遗传连锁图谱

遗传连锁图谱是通过遗传特征，描述 DNA 标记(基因和其他可以确认的 DNA 序列）的相关染色体的位点。主要研究这些标记是否一起遗传，图谱中 DNA 标记之间的距离表明了它们一起遗传的频率。这是人群遗传学的研究领域，在人类就是研究常染色体和性染色体特征的家族史。DNA 标记必须是多态性的才有用。多态性(突变)是 DNA 序列的变异，平均每 300-500bp 出现一次，代表了基因长度分布的低限。这意味着基因的多态性是非常常见的特征，如果突变发生在外显子上，尽管许多突变可以导致一些可以观察到的变化，例如眼睛的颜色、血型和疾病易感性等不同，然而突变并不一定翻译成一个改变的表型。如果突变发生在基因组非编码区，也可以作为 DNA 水平上的分子标记，但没有留下可见的表型或只能使生物体活力减弱。因为它们通常位于基因组的非编码区，因此可以被认为是隐性突变，只能在 DNA 水平识别出来。简短地说，遗传连锁图谱是在一个家族中（代代相传）观察两个标记一起遗传的频率的基础上构建的。孟德尔的豌豆颜色构成了这样的标记，尽管有些标记明显是独立遗传的（在染色体上没有相连），但有些是连锁的，他们位于同一染色体。

遗传图谱曾经用于寻找一些重要疾病基因的确切染色体定位，包括囊性纤维化、镰状细胞疾病、家族黑蒙性白痴、脆性 X 综合征和肌强直营养不良等疾病。

基因组计划的短期目标是建立高分辨率的遗传图谱 （2－5厘摩（cM，centimorgan））。两个标记如果通过重组同时被分离的几率是 1%,它们之间的距离就是 1_cM，1_cM 的距离大致相当于物理距离 1 百万碱基对（1Mb）。几年前，一些染色体的公认图谱遗传标记之间的距离是 7 到 10_cM ,最近人类遗传图谱已经达到 0.7_cM。由于重组 DNA 技术的使用，遗传图谱的分辨率提高了。这些技术包括体外放射线诱导染色体片段化和细胞融合技术（将人的细胞与其他物种细胞融合形成杂交细胞），以制备一些带有特异和多样的人染色体组分的细胞。评估放射线诱导的 DNA 片段化后的标记位点是否仍然在一起的频率，可以建立这些标记的顺序和距离。因为只需要分析染色

体的一个拷贝，所以非多态性位点在放射杂交作图中也非常有用（摘自：Primer on Molecular Genetics. Dennis Casey, Dept. of Energy，1992. http://www.bis.med.jhmi.edu/Dan/DOE/intro.html ）。

2. 物理图谱

物理图谱描述基因组或染色体中基因或 DNA 标记的分子组织。根据使用的技术不同，图谱的分辨率也大不相同。早期的方法依赖显微镜技术观察染色体致密形式的分带特征，分带通常与染色体不同的活性区域相关。在光镜下，需要在 DNA 处于相当好的组织形式时制备染色体（例如有丝分裂时）。电镜提供了更高的分辨率，可以得到更为细致的结构。

高分辨率物理图谱应用了越来越多的已知序列信息，并且将显微镜数据与遗传连锁图和 DNA 标记周围的 DNA 序列结合起来。最终的物理图谱是人基因组或染色体的全部重叠 DNA 序列。由于遗传连锁图谱是在染色体重组活性的基础上测量标记之间的距离，物理图谱和遗传连锁图谱上的相对标记间距离可以有很大的不同，这也是由于在减数分裂和有丝分裂中染色体上不同位点的重组频率不同。这种行为的机制还不清楚。它可以是独立的或与染色体结构相关的简单序列，实际上可能是由序列的特征来决定的。物理图谱和功能图谱（即标明基因功能的基因组图谱）之间的差别也非常有意思，基因组计划将为回答这些问题提供信息。

Sanger 测序中心的站点（http://www.sanger.ac.uk/）提供了人染色体的物理图谱，他们的分级结构可以将用户感兴趣的克隆定位到特定的染色体位置，并且可以逐步深入到细节直到已知的核苷酸序列。

3. 表达图谱

确认基因结构是人类基因组计划初始的驱动力；同时克隆基因在药物发现中有重要的作用。原因非常简单，因为结构基因可以被激活或灭活，所以很容易确认。实际上，确认基因的难题归

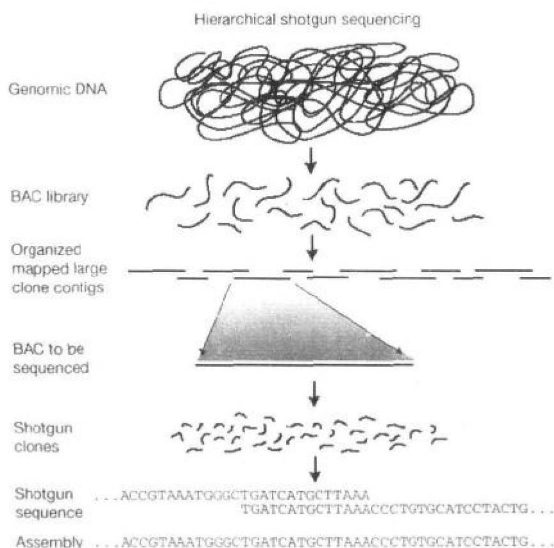


图 4-5　**公众基因组计划组织测序人类基因组的策略。首先将全基因组打碎成大的片段.克隆到 BAC 载体中,构建 BAC 文库 然后利用多态性标记排列成大片段 DNA 克隆重叠群；将每个 BAC 克隆打碎成短片段.克隆人质粒载体；短片段测序后的序列经软件装配成完整的序列。**

结为在细胞中确认 mRNA ，得到的序列标签可以作为确认新的真核生物基因的工具。这就要选择非常短的片段并测序用来构建所谓的表达图谱，而不是等待全基因组测序完成。因为基因是由编码区和包含调控序列的非编码区组成，所以表达序列标签和序列标签位点在构建人类染色体的高分辨率图谱和建立连锁时都是很重要的工具。NCBI 写到：美国人类基因组计划的一个特殊目的就是构建基因组的高分辨率 STS（sequence tagged sites ,序列标签位点，基因组中由 PCR 得到的序列）图谱。由于 EST（表达序列标签）来源于活性基因，确认 ESTs 是确认人类基因的捷径。ESTs 可以在未知任何功能时获得。由于基因并非在任何时候都表达，而

且经常以一种特异细胞类型的方式表达（也就是在生命体不同发育阶段，其表达是特异的）。所以在全部生命周期和所有生理上相关的组织中都必须检测 mRNA 的存在然后进行测序。这种方法忽略了一个真核生物基因组的大部分，但却揭示了感兴趣的生理和医学情况，这也就是功能基因组学。

非编码区 DNA通过 PCR技术逐段测序，STS 数据库的产生包括可以用于确定染色体位点的特异序列标签，可以作为基因共同分离的标记。应用电子 PCR（electric PCR），搜寻已知染色体位点的 STS 并与新序列进行比较，可以确定其染色体定位。这种方法—电子 PCR 可以用于制成各种类型的基因组图谱。

总的来说，快速测序通常导致一些结果是部分序列或未完成的序列。GenBank 的高通量分部（http://www.ncbi.nlm.nih.gov/HTGS/）试图改进这一状况，并且与提交到其他数据库的序列片段数据合作，以弥补这些部分序列或未完成序列。这里也包括日本和欧洲的数据库，这个努力的结果就是三个国际数据库的合作（DDBJ、EMBL 和 GenBank）。

4. 减少冗余性

随着基因组计划发展为一个有组织的科研项目，减少冗余性是数据库流水线化和优化过程中最受到关心的问题。冗余性的原因不仅由于不同研究者对相同的蛋白或基因感兴趣的事实，而且还由于利用不同方法随机克隆和测序基因组所产生的片段，许多片段没有生物学的相关注释。

GenBank 是一个综合性的序列数据的资源库，但选择物理图谱中的侯选序列是很困难的。这里大部分的原因是，来自同一基因的多个序列记录彼此并不相同，这些基因序列之间带有不同数量的一端序列或内含子序列。这是由于 mRNA 序列有不同的剪切方式而产生了不完整或有变异的序列。最终， ESTs 序列仅仅是片段而且有很高的错误率。在 UniGene 系统中，如果不同序列在 3’端非编码区的同源性在统计学上很有意义，这些序列就可以组成

一个 UniGene 簇。Washington 大学和 Merck 公司得到的 ESTs 序列是来自利用 Oligo（dT）为反转录引物的 mRNA ,然后定向克隆并且从 5'端和 3'端都测序，这样利用 3，端序列来组成 UniGene 簇，同时将处于同一克隆的 5 ' 端序列也纳入该 UniGene 簇（www. ncbi. nlm. nih. gov/Schuler/Papers/Esttransmap/）.

当然，冗余性对于基因组作图的某些方面和质量控制是有好处的，而且冗余性和同源性是关系非常密切的两个概念。同源实际上是指两个或多个不同基因有很大的相似性，这些序列很可能是某一人群中的等位基因或者是不同物种或生物分类中特异基因的同源基因。

三、人类基因组作图进展

在 1998 年初，人们预计人类基因组测序工作将在 2005 年完成 。1998 年十月，一个新的基因组学公司， Celera 公司(http://www/celera.com/)成立了。它是由 Perkin-Elmer 公司(http://www.perkin-elmer.com/）和 J. Craig Venter 作为发起人。在这个私人公司的竞争下，美国政府资助的人类基因组计划将“working draft (工作草图)'完成的时间已经提前到了 2001 年。目前，Celera 公司和公众基因组计划组织的工作草图已经在 2001 年 2月发表，并计划在 2003 年完成精细准确的序列图。

应用新的硬件和软件在这个发展中是非常关键的。公众基因组计划组织遵守国际放射杂交作图协议（International Radiation Hybrid Mapping Consortium），其相关信息可以通过 NCBI 获得(http://www. ncbi. nlm. nih. gov/ genemap98)。参加国际放射杂交作图协议的主要基因组中心如表 4-1。全部参加机构的名单参见 NCBI GeneMap'98.

GeneMap'98 包含了公布数据的 30,261 个基因的位点（2002 年初），已测序和定位的基因以指数级增长。位点由序列标签位点(STS) 标记并且每一个 STS 在全基因组中是唯一的。因为人类基因组只有一部分已测序，只有 3%的 STSs 序列对应实际的基因。

表 4-1　参加国际放射杂交作图协议的主要基因组中心

基因组中心	地点	互联网地址
Genethon	Evrey, France	www. genethon. fr/genethon _ en. html
The Sanger Center	Cambridge, UK	www. sanger. ac. uk/
The Stanford Human Genome Center (SHGC)	Palo Alto, CA, U. S.	www-shgc. Stanford. edu/
The Whitehead Institute for Biomedical Research	Cambridge, MA, U. S.	www. genome. wi. mit. edu/
The Wellcome Trust Centre for Human Genetics(WTCHG)	Oxford, UK	www. well. ox. ac. uk/

GeneMap'98 是从来源于表达序列的 STS 序列中得到的，代表了基因组的 3%。由于这个原因，目前的图谱实际上是指人类转录图谱。目前 NCBI 提供新的版本 GeneMap'99（http://www. ncbi. nlm. nih. gov/ genemap99），其统计数据与 GeneMap'98 相似。

GeneMap'98 和 GeneMap'99 提供了每条染色体基因的大致分布情况。由于它在 ESTs 的水平上比较了预期的基因密度和检测到的基因密度，所以关于基因的活性和基因剂量效应的相关信息可以从这里得到。很明显，沿着染色体基因分布的密度不是一致的。19 和 17 号染色体密度最高，而 18 和 X 染色体比预期值低得多，后者可能是由于女性 X 染色体灭活或男性是 XY 染色体造成的。要注意的是，目前的转录图反映了基因的表达活性，而不是编码区的实际情况，基因的密度反映一个功能分布而不是物理分布。从 ESTs 和 STSs 获得的基因密度的不同显示了基因剂量效应。

四、人类基因组序列草图公布

由国际人类基因组测序协会和美国 Celera Genomics 公司分别完成的人类基因组序列工作框架图（Working Draft）终于公布了。作为科学界的一大成就，查明人类 DNA 30亿个碱基对的序列几乎能与登陆月球、原子弹的研制相媲美。目前预测人类大约

有 32000 个基因，已经确定了其中的 22000 个，与拟南芥基因组中 25000个基因相比，数量较接近。因此，我们可以清楚看到：对于生命来说，有比基因数量更重要的东西。虽然人类基因的数量比预计的要少，但我们的基因组仍是目前被检测的最大的基因组。在已完成的基因组序列中仍有一些漏洞，一些区域需要重做。

《自然》杂志上公布的人类基因组序列是耗资 3 亿美元的成果（Nature，15 February，2001），是全世界数百名研究人员协同工作的结果。美国 Celera Genomics 公司同时也在《科学》杂志上公布了它的人类基因组序列草图（Science，16 February，2001）。公共资助的草图是全部公开的，而 Celera 的序列是有限制的。

现在人类拥有两个不尽相同的人类基因组序列草图，且存在着大量的缺口、错误、冗余以及不完整的诠释。这些问题说明每一个草图均不完美，但是许多这样的问题可以比较评估。经过对这两个草图的可比性分析，发现了一些序列的特征，包括序列缺口、连续性、这两个序列的一致性和 DNA 结合蛋白区的格式。人类基因组的两个草图分别由人类基因组计划（HGP）委员会及 Celera 遗传学公司绘制的。基因组序列详细描述了 DNA 片段的直接序列以及在这些序列重叠的基础上把小片段序列聚类成更大的单位（鸟枪聚类法）。HGP 应用的是一种不同等级作图和测序的方法，包括一系列重叠克隆的构建，这些克隆覆盖了整个基因组并对每一个克隆应用鸟枪测序法进行测序。在这些克隆上的序列重叠、作图和染色体上位置的信息的基础上，进行片段聚类分析而重建出基因组序列。Celera 遗传学公司应用的是对整个基因组进行鸟枪测序法测序，因而没有产生一系列重叠的克隆，但是也在适当的地方结合了 HGP 的信息。

第四节　功能基因组

全世界储存在数据库中的 DNA 序列有 300,000 多个，新基

因用于医学和生物学研究的潜力是巨大的。很明显在 40 多个已完成的微生物基因组计划中有大量的结构基因是新的。这意味着还没有在实验中证实它们的生物化学和生理学功能，对这些序列的结构和功能的注解（观察它们与已知蛋白的相似性）可依靠自动化的统计学分析。预测结构和功能的方法是获得其生物学信息的第一步，这些注解也越来越多基于以前预测的信息。序列的生物学含义，即表型和蛋白结构功能等，仍然是对蛋白的生物化学特性的注解。这意味着即使在基因组计划完成之后，仍然需要许多年在生理学水平上研究生物体的整体复杂性。完成全部基因组测序后首先要做的工作之一就是了解其内容，如：表型和基因型的关系。从基因组数据中提取和分析信息的任务可以通过一些公共软件实现，这些软件可以分析与 DNA、RNA 和蛋白质相关的一些特性。表 4-2 和 4-3 是一些研究基因组结构、确认新基因和它们相关的蛋白结构的常用方法以及其软件的 Internet 地址。

表 4-2 DNA 和 RNA 的公共分析软件工具

用途	软件	互联网网址
序列相似性比较	BLASTn、tBLASTx、BLASTx	www.ncbi.nlm.nih.gov/BLAST
寻找开放读框(ORF)	ORF Finder	www.ncbi.nlm.nih.gov/gorf/gorf.html
在 DNA 序列中寻找序列标签位点	Electronic PCR	www.ncbi.nlm.nih.gov/STS/
将 DNA 或 RNA 翻译为蛋白	Translate and Protein Machine	www.expasy.hcuge.ch/tools/dna.html 和 www.ebi.ac.uk/translate.html
比较基因组和蛋白质序列	GeneWise	www.sanger.ac.uk/Software/Wise2/genewiseform.shtml
寻找基因	Gene Recognition and Assembly Internet Link (GRAIL) 和 PROCRUSTES	www.compbio.ornl.gov/Grail-1.3/和 www-hto.usc.edu/software/procrustes

表 4-3　蛋白质的公共软件分析工具

用途	软件	互联网网址
序列相似性比较	BLASTp，tBLASTn	www. ncbi. nlm. nih. gov/BLAST
自动结构建模	SWISS-MODEL	www. expasy. ch/swissmod/SWISS-MODEL. html
蛋白确认和特征	Protein Identification and Characterization Programs	www. expasy. ch/tools/#proteome
寻找蛋白特征和基序	Pattern and Profile Search Programs	Expasy. hcuge. ch/tools/#pattern
结构分析	Primary Structure Analysis Secondary Structure Prediction Tertiary Structure Programs	www. expasy. ch/tools/#primary www. expasy. ch/tools/#secondary www. expasy. ch/tools/#tertiary
序列排列	Sequence Alignment Programs	www. expasy. ch/tools/#align
2维凝胶电泳分析	Melanie II	www. expasy. ch/tools/melanie/

理解基因型和表型之间的关系的第一步是观察全基因组的功能,这可以在 mRNA 的细胞表达谱中反映出来。新基因的表达与某些细胞活性相关，可以提示一些有生物学意义的信息。研究表达谱可以使我们了解未知基因的时空信息。这样一个简单的程序可以使我们按照相互关系来构建数据库。数据库可以按照蛋白和基因在不同功能水平分成亚数据库，数据库的分级结构可以按照特殊的方式帮助生物学家快速查找关于蛋白、基因、代谢途径、酶活性以及进化关系的相关信息。当前的数据库是应用相关序列、蛋白、分类信息、预测的二级结构和蛋白的结构域组织来构建的。

理解进化上有一定距离的生物（属于不同种类，如细菌和人）蛋白质家族的功能和结构是必要的。进化距离相关的生物序列经比较后可以分为三种类型：第一种就是在细胞的复制和信息储存中起作用的蛋白，它们之间有高度同源性，这些基因是真正的同源基因，它们有共同的祖先基因，形成一个蛋白家族。第二种具有相

似的结构和功能，但序列不相似。它们之间的关系可以通过结构和功能的相似性推断出来，但没有有意义的序列相似性。它们可能在进化上是或不是相关的，也 可能是会聚性进化（convergent evolution）的例子。第三种则在序列、功能或结构上都没有相似性。

因为催化位点的结构特征比全基因或全蛋白的 DNA 或氨基酸序列更保守，所以编码核苷酸结合区域的局部序列在研究进化相关性中更有价值。这一点可帮助我们查找序列中的保守特征，也正是这些特征在确定基因之间的进化关系中有意义。甚至在一些基因的全长序列与其它蛋白有很低的同源性或没有同源性时，其功能区也可能有高度的结构保守性。因为只有所选择的基因片段有同源性，这些特征可以提示一些进化机制，如基因复制或重组活动。

在加利福尼亚 La Jolla 的 Scripps 研究所（Scripps Research Institute），Adam Godzik 开发了一种新的算法来解决确认结构相似但序列不相同的蛋白的难题，基因组分析主页（http://cape6.scripps.edu/leszek/genome/）提供了 Mycoplasma genitalium、Escherichia coli 和 Helicobacter pylori 基因组与 PDB 的蛋白结构的序列比较信息，该软件将全部基因组中所有 ORFs 的预测结构与蛋白数据库中的已知的结晶衍射和核磁共振蛋白结构进行比较。比较结构中的基序（motif）可以帮助我们确认较弱的相关性（这常常在 BLAST 搜索时被忽略了），但在预测大部分细菌基因组的功能时仍然不行。将大肠杆菌基因组作为一个例子，它一共包含 4300 个基因编码 1500 个预测蛋白（全部基因或 ORF 的 40%），在这些基因之中，30%（约 500个 ORFs）不能可靠地预测其编码的蛋白结构，或不能预测其功能是什么，另外还有 30% 根本不能预测，也就是说它们是全新的蛋白。在细菌、原生质、植物或动物中没有已知的对应蛋白。

Godzik 和其他人应用的软件使用折叠预测 fold prediction）算法、特征（profile）／特征、结构／序列和结构／结构排列

(MODELLER 和 COMPOSER 软件)，这些折叠或局部结构预测使用了二级结构预测（secondary structure prediction)、埋藏氨基酸（buried amino acid）和接触特征（contact patterns）等方法。

一、未确认的读框（Unidentified Reading Frames，URFs)

基因组计划完成后有 30%-40%全新而且未确认的基因序列，这些指的是 URFs（未确认的读框),它们没有相关的生物学信息。这些序列没有已知的同源基因，所以它们肯定会编码出一些功能还没有被生化学家或微生物学家发现或研究的新蛋白。一些如 Adam Godzik 等的结构预测算法在这里有很大的帮助，但对了解其功能则没有帮助。这说明结构和功能关系之间的知识还有许多未知的东西。通常情况下，这里面也有一个似乎基于实验证据上的文饰心理。许多预测结构的方法是统计学的方法，它依赖从已知的结构中获得的信息。有限的样品数量（已知结构的数量）限制了预测工具的准确性。

研究新基因进化关系的另一个可选择的工具是通过研究基因组“行为”直接进行基因组比较。一种生物已知蛋白的突变率是多少？这个信息可能对预测一种生物中的一个新基因的唯一性有帮助，该基因与任何已知的蛋白都没有同源性。假设一种生物的全基因组突变率是平均的，一个 URF 序列的不同点暗示着它可能属于蛋白的一个新类型。

将催化连接氨基酸与某些转运 RNA（tRNA）的酶作为例子（见 F. Doolittle，1998，Nature 392：339)，在哺乳动物中至少有 20 种不同的酶，每一种对应一种氨基酸用于合成蛋白。在已完成的原生质 M. jannaschii 基因组计划中，有四种乙酰化氨基酸的 tRNA 合成酶的相应基因没有确认出来，尽管它们肯定存在。因为所有的 tRNAs 与其相应的氨基酸是适当连接的。缺乏该基因的一个可能的解释，就是假设乙酰化氨基酸 tRNA 合成有一个全新的机制：对连接到 tRNA 分子上的氨基酸的化学修饰。有一个未确认基因编码的赖氨酰 tRNA 合成酶，这是在一个与基因组计划无

关的研究工作中证实的，这种合成酶负责将赖氨酸和其相应的tRNA 连接，它的序列与任何已知的赖氨酸-tRNA 合成酶序列不同。实际上，这就是全新蛋白质家族的一个例子，结论是从 DNA 序列可以判断完全不相关的两个蛋白质可能执行相同的酶活性。这种现象可以在其他种类的酶中见到，例如丝氨酸蛋白酶、糜蛋白酶和枯草杆菌蛋白酶。尽管有不同的底物，它们利用结构保守的活性区域催化相同的化学反应。

蛋白质在进化上不相关而功能却相似的事实，证实了在完全不知道功能数据的情况下，推测 DNA 序列能编码出何种蛋白质是很困难的，有时甚至是不可能的。应用来自功能研究的数据对于确认 URFs 的生物学功能是必需的。为了理解乙酰化氨基酸-tRNA 合成酶结构和功能之间的关系，需要精通这方面的知识，不熟悉 tRNA 代谢的科学家不可能发现这里明显的关系。

二、同源异种组（Cluster of Orthologous Groups：COGs）

在分类不同的生物中（同源异种组， orthologs）发现基因之间的关系，与在同一种生物或同一群体中发现基因的关系同样都是基因组计划真正的潜力。经过比较 8 种已完成的基因组编码的蛋白质序列，它们代表了 6 个主要的进化树分支（http://www.ncbi.nlm.nih.gov/COG；1999 年 4 月）。在 NCBI 的 COGs 主页上显示了同源异种组（COGs），目前这一计划已经包含了 44种已完成的基因组，共有 3311 个 COGs（2002 年 1 月）。这是一种应用数据库通过联系不同已完成基因组序列信息来产生新信息的尝试。按照 NCBI 构建 COGs 所采用的功能定义，任何两个蛋白质如果来自属于同一 COGs 的两个不同物种，它们就是同源异种组，而且假设是通过物种进化形成的。同源异种组也包括同源同种组，来源于基因复制活动。

应用已完成的 8 种生物基因组（E. coli；H. influenzae；M. genitalium；H. pylori；M. pneumoniae；Synechocystis；M. jannaschii；S. cerevisiae 等），一共确认了 864 个 COGs。它们分别

属于信息储存和处理组——J、K、L 组，细胞合成组——O、M、N、P 组，新陈代谢组——C、G、E、F、H、I 组和预测或未知功能组——R、S 组。后者是未确认功能的一个组，一共包括 180个 COGs，包含与预测功能相关的 1828个蛋白质和结构域（R）以及 271 个没有特征的蛋白质或结构域（S）。通过对 COGs 的分析可以使我们理解进化的关系，并且在物种分类之间确认相关的功能。

N 组包含 20个 COGs，其中之一代表了在真菌和真核生物起作用的蛋白酶一信号肽酶家族（COG ID0681），但在原生质中是没有的。信号肽酶是一个小的膜结合蛋白，负责在真核生物内质网以及细菌和线粒体内膜转运蛋白时切除蛋白质的 N 末端信号肽。信号肽酶 I COG 包含 8 个成员，1 个 E. coli 蛋白 LepB，H. influenzae 蛋白 HIN1152，Synechocysis sp. 同源同组物（paralogs）alr377 和 sll0716，M. jannaschii 蛋白 MJ0260 和 3 个酵母同源同组物 YMR150c、YMR035w 和 YIR022w。

COG 证实了在进化距离上相关生物的蛋白质之间寻找进化树关系的复杂性。酵母同源同组物之一 YMR035w 与三个不同的菌种（E. coli、H. influenzae 和 Synechocystis sp.）显示了很好的相似性。YMR150c 和 YMR035w 是酵母线粒体内膜上的蛋白酶，负责在内膜上切除某些蛋白质的信号肽，但有不同的底物特异性。YIR022w 是酵母内质网上的信号肽处理蛋白质，在信号肽切除和正常蛋白分泌速度中起作用。

各簇的树状图显示了线粒体蛋白酶与细菌同源异组物的相近关系以及内质网蛋白酶同源同组物与原生质 M. Jannaschii 同源异组物之间的关系，这与真菌都有一个共同的单细胞生物祖先的理论是一致的。M. jannaschii 蛋白酶与酵母和 cyanobacteria 都有同源异组关系，这强调了在分类学上原生质与真核生物和真菌是不同的。这个 COG 分析说明，在三个酵母肽酶中两个线粒体亚型是真正的同源同组物并有共同的真菌起源，而内质网亚型是独立进化的或起源于一个较古老的祖先基因。

COG 数据库列出了相关基因的特征，这些特征表明它们在不同的生物中出现。其中一个特征是 eh-cmy，这个特征排除了两种革兰氏阳性致病分支疟原虫和导致溃疡病的 H. pylori，但包括了革兰氏阴性致病源 H. influenzae。具有这个特征的 39 个其他的 COGs 也被确认了，其中包括信号肽酶 I。COG 代表了最共同的进化遗传特征，在 NCBI 的 8 种基因组分析中包含了 110 个簇，其中大部分属于与翻译、核糖体结构和生物起源相关的 J 功能组。其他酶的功能组属于一些中心代谢途径，如糖酵解、戊糖磷酸化途径、RNA 聚合酶、蛋白折叠和分泌等。在不同生物的酶和代谢途径中确认这些特征可以得到一些有关生命在不同环境下生存的必要条件的生物化学信息。

最适合用于研究感染宿主和复制所需要的最少遗传物质的致病原组是病毒。由于病毒应用宿主生物细胞内的功能，所以它们并非复杂系统。它们的基因组很明显是由一些适应性的必要基因组成的。病毒的适应性是最好的，但又是独立的包含最小基因组的最小生物。由于病毒基因组非常小，所以在完成第一个致病菌 H. influenzae 测序很早以前，病毒的基因组就被测序了。噬菌体 ΦX174 的基因组是第一个完成测序的病毒基因组。其 DNA 序列包含 48,502 个碱基，由 Frederick Sanger 小组于 1982 年完成，细菌 H. influenzae 基因组测序在 1995 年完成，包含一百七十万个碱基。

第五节　人类基因组计划与生物信息学研究

人类基因组计划（HGP）目的之一，就是找到人类基因组中的所有基因。除功能克隆和定位克隆等策略之外，生物信息学为分子生物学家提供了一条寻找和研究新基因的新思路，即从高度自动化的实验出发，经过数据的获取与处理、序列片段的拼接、可能基因的寻找、基因功能的预测一直到基因的分子进化研究。这

个过程的每一个环节，都是生物信息学研究的重要内容。

一、高度自动化的实验数据获得、加工和整理

如何将实验室中得到的生物学信息转化为计算机能够处理的数字信息，是生物信息学的一个重要课题。这种转化更多地体现在各种自动化分子生物学仪器的使用上，如 DNA 测序仪、PCR 仪等。这类仪器可将实验所得的物理化学信号转化为数字信息，并对其作简单的分析，再将分析结果用于实验条件的控制，完成高度自动化的实验过程。从事大规模 EST 测序和 DNA 物理图谱构建的实验室都已建立起高度自动化的机器人系统来完成大部分的实验工作。

伴随着实验过程的高度自动化甚至工厂化，从事大规模分子生物学项目的实验室，每天需要存储的数据可以轻易地超过几千兆字节。这样大的数据量必须用专门的实验室数据管理系统进行处理，以自动完成包括实验进程和数据的记录、常规数据分析、数据质量检测和问题的自动查找、常规的数据说明和数据输入数据库在内的各项工作。由于不同实验室需处理的数据类型各不相同，目前各个实验室都是各自开发自己的系统，还没有成熟的可用于不同实验室的分子生物学数据管理系统。但随着测序逐渐成为实验室的常规工作，对这种系统的需求会越来越大，此类系统的发展将成为大势所趋。

二、序列片段的拼接

目前，DNA 自动测序仪每个反应只能测序 500bp 左右。如何将这些序列片段拼接成完整的 DNA 顺序就成为接下来的一个重要工作。传统的测序技术通常将克隆进行亚克隆并对亚克隆进行排序。这些工作需要大量的人力物力。现在生物信息学提供了自动而高速地拼接序列的算法，即根据 Lander-Waterman 模型利用鸟枪法进行测序，再将大量随机测序的片段用计算机进行自动拼接。这种技术不仅避免了亚克隆排序所需的大量繁琐的工作，还使序列具有一定的冗余性以保证序列中每个碱基的准确性。

序列拼接算法的进一步发展，需要在以下方面进行改进：(1) 将已知的基因组知识应用于拼接算法，以进一步提高拼接真核基因组的有效性。(2) 自动处理自动测序造成的差错，特别是对有差错倾向的 EST 序列更是如此。

三、基因区域的预测

在完成序列的拼接后，我们得到的是很长的 DNA 序列，甚至可能是整个基因组的序列。这些序列中包含着许多未知的基因，下一步就是将基因区域从这些长序列中找出来。

所谓基因区域的预测，一般是指预测 DNA 顺序中编码蛋白质的部分，即外显子部分。不过目前基因区域的预测已从单纯外显子预测发展到整个基因结构的预测。这些预测综合各种外显子预测的算法和人们对基因结构信号（如 TATA box 和加尾信号）的认识，预测出可能的完整基因。

在介绍算法之前，我们先介绍一下衡量一个算法优劣的标准：敏感性（sensitivity）和特异性（specificity）。假设待测序列中有 M1 条序列是基因序列，剩余的 M2 条为非基因序列。我们用程序对待测序列进行预测，N 条序列被预测为基因，其中有 N1 条确实为基因，其余 N2 条不是基因的一部分。敏感性定义为 N1/M1，它表示程序预测的能力。其特异性定义为 N1/N，它表示预测结果的可信度。敏感性和特异性往往是一对矛盾，一般以敏感性和特异性的平均值作为评判程序优劣的标准。

预测外显子的基本算法，早期有最长 ORF（open reading frame）法。在细菌基因组中，蛋白质编码基因从起始密码 ATG 到终止密码平均有 1000bp，而长于 300bp 的 ORF 平均每 36kb 才出现一次。所以只要找出序列中最长的 ORF（> 300bp）就能相当准确地预测出基因。核苷酸语汇（nucleotide words，即数个连续核苷酸的排列）选用频率的统计差异也被用来区别编码和非编码区域。这种差异可能来自编码和非编码区密码子选用的差异和周期特征的差异，其中一个显著的特征是 6 核苷酸的选用差异。

在目前的各种预测程序中这是一种被广泛应用的方法。近年来同源比较算法也被应用于预测可能的基因。许多基因预测的程序都已经整合了同源比较算法，比如著名的 GRAIL 程序。

除上述提到的算法之外，目前被应用于基因预测的算法还有：法则系统（rule-based system）；语言学(linguistic)系统；线性判别分析(Linear Discriminant Analysis，LDA）；决策树(decision tree)；spliced alignment 算法；傅利叶分析(Fourier analysis）等。

综合以上算法和人们对基因结构信号知识的基因预测程序已有不少。其中有的对编码序列的预测准确率高达 90% 以上，并且在敏感性和特异性之间取得了很好的平衡。

四、基因功能预测

用实验手段证实一个预测的新基因后，下一步要做的就是寻找这个基因的功能。生物信息学为此提供了一系列方法，使我们的研究能够有的放矢。

1. 序列同源比较

序列同源比较往往是得到新基因后预测其功能的第一步。通过同源比较来预测基因功能是基于这样一个假设：如果基因 A 与基因 B 有相当的同源性，那么基因 A 可能具有类似基因 B 的功能。利用同源比较算法，将待检测的新基因序列在 DNA 和蛋白质序列数据库中进行同源搜索后，我们可以得到一系列与新基因同源性较高的基因或片段。这些基因和片段的已知的功能信息就为进一步研究新基因功能提供了具有相当参考价值的导向。

2. 同源比较的发展方向

用于将序列在序列数据库中进行同源比较的 3 种流行的算法：Smith-Waterman 算法、FASTA 和 BLAST 算法。它们有各自的优缺点。面对飞速增加的数据库数据，如何同时获得高敏感性和高速度仍然是一个课题。

同源比较算法中另一个需要继续发展的方面是同源比较算法中使用的计分矩阵的完善，特别是间隔的计分方法的研究。研究

证明，使用更好的计分矩阵能够使算法的敏感性显著提高。

需要解决的另一个问题是目前数据库中部分数据的冗余度太高。特别是 EST 数据库，某些基因甚至有数千条 EST 与之对应。所以对数据库进行同源检索所得到的结果可能是一大堆无用的信息淹没了有用的信息。这个问题可以通过屏蔽掉检索序列中的重复顺序，或清除数据库中冗余数据的方法得到部分的解决。

3. 寻找蛋白质家族保守顺序

通过同源检索，我们可能推测待检的新基因是某个蛋白质家族的新成员，下一步就是寻找新基因中包含的该蛋白质家族的保守序列。这样，也就为进一步深入地研究其功能作好了准备。

多序列同源比较，或称为多序列对齐（multiple-sequence alignment），是将多个序列进行同源比较以发现其共同的结构特征的方法，被广泛用来寻找基因家族或蛋白质家族中的保守部分。Feng-Doolittle 算法是较常用的多序列对齐算法。其他的新算法包括 HMM 方法，Gibbs sampling 以及处理多结构域蛋白质家族的算法。由于保守部分往往与家族成员的功能密切相关，所以通过这些方法建立蛋白质家族数据库，能够帮助科学家更好地认识基因的功能。这些数据库可以帮助我们把新基因所属的蛋白质家族及其保守部分找出来，并提供这个家族其他成员的结构和功能信息。

4. 蛋白质结构的预测

有时一个可能的新基因通过同源检索找不到任何同源基因。这种序列就称为“孤儿”基因。生物信息学也提供一些预测孤儿基因功能的方法。这就是通过基于结构的同源比较（structure-structure alignment）寻找结构同源的基因或直接预测其高级结构来推测其可能的功能。有许多蛋白质高级结构数据库提供结构同源比较的检索。另一方面，直接预测基因产物的高级结构的算法现在已经有不少，然而，由于蛋白质的折叠结构实在太复杂，使得计算最佳构象非常困难。如果结构生物学在这方面的研究能够

有所突破，无疑将大大推动基因功能的预测。

5. 分子进化的研究

通过上述种种方法我们可以预测出一个新基因可能具有的功能。然而预测新基因只是生物信息学研究的一个方面，这门学科的根本目标是探究隐藏在生物数据后面的生物学知识。对于基因组研究来说，一个重要的研究方向就是分子序列的进化。通过比较不同生物基因组中各种结构成分的异同，可以大大加深我们对生物进化的认识。这些研究已逐步形成一个称为比较基因组学的新学科。从各种基因结构与成分的进化，密码子使用的进化，到进化树的构建，各种理论上和实验上的课题都等待生物信息学家的研究。

参考文献：

1. Altschul SF, et al. Basic local alignment search tool. J Mol Biol, 1990, 215 (3): 403-10
2. Sali A, Overington JP. Derivation of rules for comparative protein modeling from a database of protein structure alignment. Protein Sci, 1994, 3 (9): 1582-96
3. Topham CM, et al. An assessment of COMPOSER: a rule-based approach to modelling protein stucture. Biochem Soc Symp, 1990, 57: 1-9
4. Sanger F, et al. Nucleotide sequence of bacteriophage lamda DNA. J Mol Biol, 1982, 162 (4): 729-79
5. Venter JC, et al. The Sequence of the Human Genome. Science, 2001, 291 (5507): 1304-51
6. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature, 2001, 409: 860-921

第五章　蛋白质组分析

一般认为上世纪 90 年代是基因组的十年，而新世纪的头十年将成为蛋白质组学快速发展的十年。利用蛋白质组学技术生成的定量表达数据第一次可以在规模和敏感性上与基因水平相媲美。这个进展对于我们理解人类健康和疾病的细胞组成结构以及对药物、农业和生物技术等有着重要的意义。确实，蛋白质组学业已在大范围应用中产生了重要发现。

随着双向电泳、质谱以及不同研究方法核心技术的改进，蛋白质组学保持着持续快速的进步。完全注释的蛋白质组学数据库现在在一些领域已经出现了，并为系统研究提供了一个平台，在临床应用诸如心血管和肿瘤学中尤其有发展前途。在蛋白质水平上的大规模定量研究也正变为现实。

第一节　蛋白质组学

大部分数据库由基因、基因组和蛋白质序列组成。在单细胞生物（如细菌和酵母）或多细胞生物（如植物和动物）中，核酸、蛋白、脂类和碳水化合物这些大分子作为细胞的结构并行使功能，但如何在一起起作用的机制还不清楚。而且 mRNA 是蛋白质生物合成的中介分子，mRNA 的水平代表基因表达水平。因此，mRNA 序列被用于建立 EST（expressed sequence tags，表达序列标签）数据库。然而，细胞 mRNA 表达量并非蛋白质表达量的可靠标志。因此，建立细胞中蛋白表达谱至关重要。这不仅可以得到蛋白表达的相对数量，而且可以得到其存在形式。翻译后的修饰过程如糖基化、酰基化、辅基、磷酸化或水解等影响蛋白质活性，所有

这些过程都可以控制细胞中蛋白质的活性和定位。

更加复杂的是，在细胞周期的不同时段、细胞代谢和环境应激中、细胞间信号传递或疾病个体中，细胞的蛋白质组成和翻译后修饰都是有所不同的。例如肿瘤与正常组织相比，关键蛋白质的表达和活性会有变化。由于这些蛋白质与生长控制、肿瘤发生有关，正常时应当停止分裂或程序性死亡（老化或凋亡）的细胞表现为失去生长控制或无限制倍增。

所有体细胞都包含了全部基因组，但只使用其中一部分发挥其调节活性。基因的上调和下调对细胞的发育非常重要，可以使单个细胞倍增并分化为特殊的细胞类型、组织和器官。基因的活性谱实际上说明了细胞控制基因产物水平，并使之表现为隐性或显性基因的过程。这被称为基因剂量效应。X 染色体相关的基因在男性只有一个拷贝而在女性则为双拷贝，因此有很多的研究报道。蛋白质的表达水平通常影响代谢和第二信使的途径，因此经常被用作微调控制系统。基因剂量效应的细胞机制还有待进一步研究。

并非所有的大分子都依赖线性模板来合成，实际上只有 RNA 和蛋白质这两种分子的合成是由 DNA 编码的，而其它所有的细胞内活动（包括前面提到的蛋白修饰）都是通过分子间相互作用、时序合成（Sequential synthesis）和空间分隔（区室化）来完成的。不依赖基因模板的一个很好的大分子例子是多糖和蛋白质、脂类的糖基化。多糖（碳水化合物）以线性形式存在，也可以是分支或多聚体形式，多聚体顺序由细胞内机制（蛋白质催化碳水化合物合成）催化形成，这些顺序并不是象 DNA 编码蛋白质合成那样由其他线性分子模板来编码。相反，多糖的合成是一个顺序的催化过程，是由细胞内酶的空间排列来完成的。多糖合成过程中缺少了某种蛋白质可以使这个过程发生障碍，因此合成多糖的一组酶的作用并不是独立的。

了解酶作用途径的结构组成非常重要。对于新发现的基因可以在不同种属之间比较单个基因的同源性，而且可以比较其在细

胞整体中所起的作用但有一些问题，如：是否在不同种属间所有酶作用途径都相同？在相同的酶作用途径中是否所有的酶都有同源性？某一种酶在作用途径中是否比其他酶更重要或更保守？是否在某些种属中存在一些与其它种属不同的酶作用途径？这些问题对二十一世纪的生物学是真正的挑战。国际互联网（或同样形式的公共交流工具）在这一发现过程中是重要的工具，它可以提供一些数据库，用于比较细胞或生物从发病到死亡这一阶段中的蛋白组成变化、代谢活性以及功能变化。

分析在特定状态下某一细胞类型或生物所有的蛋白质，这是蛋白质组学（Proteomics）这门新学科的任务。蛋白质组学的目的是分析不同时期细胞或生物体蛋白质的组成或表达情况，最终阐明细胞体进行代谢、信号传导和网络调控的组织方式和动力学。它的基础是阐明细胞内复杂的酶作用机制。蛋白质组学这个词的含义是指对生物体中决定某一状态的所有的蛋白质之间相互作用的研究。明确这些相互作用对于了解其生物学信息包括进化轨迹非常重要。

一、蛋白质组学研究的策略和技术

蛋白质组学的研究涉及到的特别技术主要有 2D 电泳和质谱分析。2D 电泳是将蛋白在多聚体凝胶中进行二维电泳，其中一个方向是蛋白质分子量，另一个方向是与 pH 值相关的蛋白质所带电荷量，该技术称为二维凝胶电泳（2-D 胶）。它可以用于比较生物生命周期中不同时期的蛋白质表达量及其修饰情况，它可以比较一组蛋白与另一组蛋白之间的差异。二维凝胶电泳的优点是不仅可以代替研究单个蛋白的费时费力的工作，而且可以研究同一时间表达的蛋白之间的相互间关系，进一步可以把这种相互关系与细胞活性联系起来。

由于许多蛋白在二维胶中的特性还不了解（或不十分了解），蛋白质组学仍然是一个艰难的工程。生物化学家最耗费时间的工作是准确地确认二维胶上每一个点的蛋白性质、是否发生修饰以

及是否只是片段。正如前述，二维胶显示蛋白质的两部分信息，分子量大小和电荷。这两个物理参数值取决于细胞在蛋白分离和纯化那一时刻的状态。利用 DNA 序列计算得到的蛋白质分子量并不是总能与电泳得到的分子量相符，因为电泳得到的分子量反映的是蛋白在凝胶基质中的可溶性和泳动率等综合情况。在分子量方向上反映的泳动率不仅依赖蛋白质真正的分子量，而且更能准确地反映出蛋白质所带电荷量。改变电泳系统的 pH 值可以引起蛋白质泳动率的变化，这是由于改变了蛋白质的电荷／分子量的比值，并非所有分子量相同的蛋白质都具有相同的电荷／分子量比值。精确地确认凝胶上蛋白质分子量和序列对于解释电泳结果非常重要。

蛋白质组学研究中应用的另一项重要技术是质谱分析。质谱技术的原理在于产生不同大小的气化的样品离子，并根据不同离子间的质量／电荷比的差异来确定分子量及分子结构。通常质谱仪由进样装置、离子化源、质量分析器、离子检测器和数据分析系统等组成。传统的质谱仪通过加温蒸发方式将小分子物质离子化，只能检测相对分子质量在几千道尔顿左右的分子，对生物大分子、非挥发性物质和热不稳定物质则无法检测。近年来，以电喷雾电离（electrospray ionization，ESI）和基质辅助激光解析离子化（matrix-assisted laser desorption ionization mass spectrometry，MALDI-MS）为代表的软电离技术的发展，将质谱的灵敏度和高质量检测范围提高到 fmol（10^{-15}mol）以上的水平，检测的相对分子量可高达几十万道尔顿的生物大分子。从而开拓了质谱学一个崭新的领域—生物质谱。

电喷雾电离（ESI）利用位于毛细管和质谱仪进口间的电势差，直接从液相物质生成单价或多价离子。样品在电场的作用下成为以喷雾形式存在的带电液滴，并通过干燥气体或加热，使溶剂蒸发，最后形成气相离子，然后通过质量分析器分析离子的质量／电荷比。ESI 的优势在于可以方便地与分离技术（如：毛细管

区带电泳）联用。MALDI-MS 以有紫外吸收的小分子晶体作为基质，样品与基质结合后，用特定波长的激光照射基质会迅速升温，样品和基质迅速气化。在气态下样品与基质反应而离子化，在静电作用下直接引入质量分析器。 MALDI-MS 最大的优点在于可以直接分析混合物，这是由于 MALDI 主要生成单电荷离子，质谱图中的谱带与混合物中不同蛋白质和基质之间存在对应关系，适用于糖蛋白和蛋白酶消化产物的分析，在蛋白质组研究中特别有用。

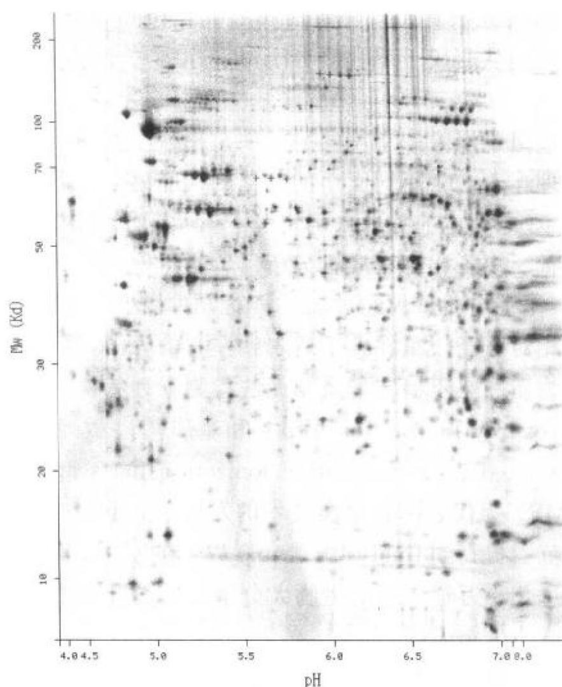


图 5-1　人红白血病细胞的二维凝胶电泳图（来自 http：//www.expasy.ch/ cgi-bin/map2/），点击其中标记红色的点可以显示该点代表的蛋白信息，并显示它在其他细胞或组织二维凝胶电泳图的位置。X 轴为 pH 值，Y 轴为分子量

现代自动分析系统有助于大规模地确认 2D 凝胶中的蛋白点（图 5-1）。目的蛋白可以在凝胶介质中消化，提取得到的多肽并加入高质量精确的 MALDI-MS （基质辅助的激光解析离子化质谱）来分析。这里肽段被离子化并可以确定其电荷／质量比，将质量／电荷比率与可能的氨基酸序列相配对。如果配对结果模棱两可，则将肽段进行微测序，并用 BLAST 工具与数据库进行同源性比较。如果从一个二维凝胶点中得到的许多肽段都与数据库（如 GenBank）中的一个序列相配，那么就可以成功地确认相应于这个点的蛋白质了（图 5-2）。

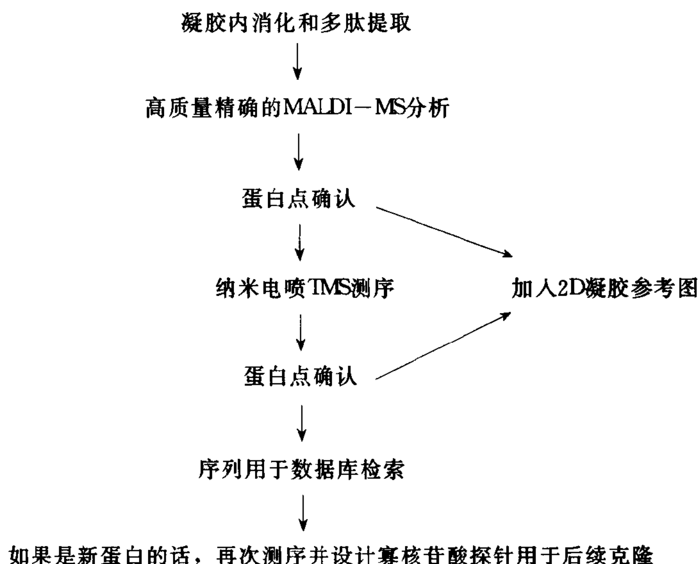


图 5-2　从二维凝胶电泳中确认蛋白质的策略

由于肽段有可能被化学修饰，所以配对并非容易作到。细胞中的这些翻译后修饰通常控制着蛋白的活性。这些修饰可影响蛋白质的净电荷、活性和可溶性。磷酸化这种修饰方式可以给蛋白

加入负电荷，进而影响了电泳中的泳动率。单个负电荷的引入对泳动率的影响相当于减少 2KD 的蛋白质分子量，大约相当于减少 15-18 个不带电荷的氨基酸。糖基化也影响蛋白质的分子量，但并不一定受 pH 值的影响。因为有多种修饰以相似的方式来影响蛋白的泳动性，所以在二维凝胶中分析蛋白质细小的泳动性差异并非易事，这需要细致的生物化学分析。

在近几年，肽段确认的全过程已经实现了自动化。自动化过程需要特殊的机器人设备以及专业软件。计算机在自动化控制和分析过程中扮演着中心角色。一个自动收集器能收集从高压液相色谱中按分子量大小分离的肽段样品，极少量样品加入毛细管柱中，经过纳米电喷离子化后用于质谱分析。分别用于实验和预测的质谱，可产生相关的交叉数据以确认提取肽段的序列。如果来自一个二维凝胶点的几个片段都与数据库中的某个氨基酸序列相符，一个蛋白就可以确认了。

二、EXPASY 的二维聚丙烯酰胺凝胶电泳数据库

细胞的蛋白质组分析的第一步是比较经激活剂刺激细胞后（例如用胰岛素作用于肝细胞）和未受刺激的细胞提取物的二维电泳图。许多公共数据库包含了一些初步确认的新蛋白质的分子量和电荷的二维凝胶信息，并且新蛋白质的数量逐渐增多。公共蛋白质组数据库 SWISS-2DPAGE（http://www.expasy.ch ）由瑞士日内瓦大学医院建立。该数据库的目的之一是在功能水平上通过直接研究基因产物和相应的翻译后修饰来理解生命体的相关机制。该站点可以通过交互方式获得二维凝胶数据库，提供在线帮助和二维凝胶电泳的技术手册。送去的样品可获得二维凝胶电泳服务，另外还提供培训课程（非在线服务）和分析二维凝胶电泳结果的软件包。

Expasy 的二维凝胶电泳数据库包含了多种组织和生物的电泳信息，目前公布的版本 SWISS-2DPAGE Release15.0 中包含了来自 33 个参考图谱的 861 个条目。来源包括酵母、大肠杆菌、

Dictyostelium、拟南芥、以及多种类型的人类组织和细胞（血小板、红细胞、巨噬细胞、血浆蛋白、淋巴瘤、肝、肾、两种白血病细胞系、脑脊液、结肠上皮细胞、结肠腺癌细胞系（DL-1）、HepG2蛋白和 HepG2 分泌蛋白等），还有多种类型的小鼠组织。已知蛋白可以用 SWISS-PROT 的序列号查询，还可以直接点击二维凝胶图中标记的点来查询。如果新蛋白的氨基酸序列已知，其预测的电泳位置可以确定。假定的蛋白分子量和电荷可以帮助我们在凝胶中定位蛋白质，然而蛋白质在凝胶中的溶解性和氨基酸的翻译后修饰经常使其理论值和实验值不同，这一点要在二维凝胶电泳分析中重点强调。由于上述原因，许多蛋白在二维凝胶电泳中可以产生多个点，这个信息对于生化学家分析细胞内环境中蛋白质的功能非常重要。

经过研究，人们发现电泳图谱中大部分位点与任何已知蛋白并不相关。在二维凝胶中更快地确认蛋白质的新技术正在开发之中。用于肽段微测序和质谱研究的生化分析手段与核酸文库的测序方法非常相似。

一旦某种细胞或生物的某种蛋白质被确认，就可以与其它种类的细胞或组织进行比较，也就有可能揭示相互间不同表达水平或翻译后不同的修饰方式。当然，用这种方法比较蛋白质的表达水平并不是二维凝胶电泳很重要的一项功能。蛋白质在凝胶中的泳动行为主要依赖其纯化方法、来源和电泳程序。对各个位点相对位置和强度的细致描述可用来比较不同的电泳结果。SWISS-2DPAGE 提供了一个分析软件包，它可以进行快速图象处理和全二维分析，为全世界的研究者提供参考；它能完成自动二维凝胶配对和比较。(Melanie II 2D 分析软件，由日内瓦 Melanie 小组的 Denis Hochstrasser 开发，http://www.expasy.ch/melanie/MelanieII/description.html)。以下是 Melanie II 的特征：

快速图象处理：

- 图象放大缩小

- 滤镜功能（图象平滑、对比度增强、背景消减）
- 凝胶翻转
- 凝胶堆积（以利于图象更好地显示）
- 图象伸展

全二维凝胶分析：

- 自动点确认和分析
- 高斯点建模
- 凝胶覆盖显示
- 即点即现的界面
- 内建的标尺
- 等电点／分子量设定
- 丰富的报告
- 直方图
- 统计数据分析

全球的电泳图比较：

- 多个凝胶显示
- 快速、自动凝胶比较和配对
- 对其它凝胶比较提供参考凝胶数据
- 合并一套凝胶产生综合凝胶
- 用 SWISS-2DPAGE 来管理凝胶数据
- 通过 Expasy 网络提供与其他生物学数据库的超链接，如：SWISS-2DPAGE 和 SWISS-PROT 数据库。
- World Wide Web 服务器

数据输入／输出：

- 凝胶打印
- 从 TIFF 或 PPM 图象输入／输出
- 数据输出到 Excel 或其它应用程序
- 数据以 Melanie I 格式输出到公共统计和探索分类程序。

三、其它的二维凝胶电泳数据库

有两个特殊蛋白质组数据库，可用来比较与毒素和异种生物相关的蛋白质表达谱：1、位于 Oxford's Glycosciences 的啮齿类动物分子效应数据库（Rodent Molecular Effects Database）；2、位于 Human Genome Research 丹麦中心的角化细胞数据库（http://biobase.dk/cgi-bin/celis）。后者提供基因敲除（knockout）和转基因动物的数据，基因敲除和转基因动物的含义就是使特异基因失活或将基因加入到动物胚胎干细胞中。基因功能的失活和表达需要在二维凝胶电泳中观察其蛋白质水平上特异信号的缺失或出现。例如，美国 ESA公司的神经系统疾病数据库就是通过蛋白质差异显示的方法来研究阿尔茨海默病、帕金森氏病和 Huntington's 病。

酵母蛋白组学数据库（Yeast Proteome Database，YPD）是蛋白质组公司（Proteome Inc.）（www.proteome.com）建立的，这是企业涉足蛋白质组学研究的例子。它将现存的大量文献组合成一种特殊的形式，包含有啤酒酵母（Saccharomyces cerevisiae）的所有蛋白质数据，啤酒酵母的基因组测序已于 1997 年完成。

YPD 是关于酵母已知蛋白和酵母基因组计划预测蛋白信息的百科全书。这些信息与蛋白质的基本生物物理和功能数据相关，包括：用质谱分析得到的蛋白质分子量、从基因组序列中预测的氨基酸序列以及文献中报道的蛋白质功能等。目前，数据库中已有约 30 个新的酵母蛋白质的信息特征在不同水平上得到明确，和 3000 多个部分信息特征（主要指含有多少 ORFs 和 URFs）明确的蛋白质。另外，还包括了一些同源性的信息。例如，酵母与人类蛋白的同源性等。把酵母作为模式生物来研究在人类代谢、生理过程中与酵母作用相似的蛋白质，所得到信息非常重要。YPD 收录的数据和二维凝胶图总体上包括了来自质谱分析的分子量信息、来自氨基酸序列的电荷和化学修饰信息以及大部分文献报道的功能信息。

蛋白质组公司还提供了人类蛋白质组监测数据库

HumanPSD、G 蛋白偶联受体蛋白质组数据库 GPCR-PD、线虫 C. elegans 蛋白质组数据库 WormPD、酵母 S. pombe 蛋白质组数据库 PombePD 和人类致病原真菌数据库 MycoPathPD。

第二节　代谢通路的重建

一、京都基因、基因组百科全书—KEGG(Kyoto Encyclopedia of Genes and Genomes)

建立京都基因、基因组百科全书（KEGG）的目的之一是试图提供一种利用计算机模拟细胞中分子信号途径的方法。KEGG 是日本京都大学化学研究所日本人类基因组计划的一个组成部分(http://www.genome.ad.jp:80/kegg/)。KEGG 的基础结构与 NCBI 网站相同，因此它面临的技术挑战也非常大。KEGG 收录有关代谢过程中的分子间相互作用的信息，其目的是有利于寻找现代分子生物学一些普遍问题的答案，诸如基因序列和蛋白质功能之间的关系、细胞内蛋白质折叠问题、功能重建的难题、基因组和生物中有关发育和形态的问题。KEGG 的目标是以各种分子中不同成分之间的关系为基础建立一个功能图谱，这个功能图谱显示出各种代谢和调节途径，来源于基因组图谱、碱基序列、基因的物理图和遗传图以及 LIGAND 数据库（包括酶、复合物及其组分）的各种分子。

二、功能重建模型（The Functional Reconstruction Model）

有一些数据库和研究机构也含有与 KEGG 相似的信息，但后者有一点显著的不同之处，KEGG 还包括一个推断数据库。KEGG 的用户可以利用基因或分子之间的布线图来估计分子间的转导途径和双方的关系。将细胞理解为一个复杂的、可以自身装配的整体，这样利于更好地理解细胞中不同组分之间的关系，这就是功能重建模型。正如 KEGG 里描述的那样“基因组只是仓库的一部分，而基因组中所有的调节信号也只是编码调节的一个小

部分。以这个观点来看，生命的蓝本（Blueprint)应该将细胞作为分子间相互作用网络的整体来书写。”为了更好地理解这个分子间相互作用的网络，KEGG 应用一种预测工具来研究这个“仓库”中单独组分之间的新的相互关系。这些工具放在 KEGG 的“Search and compute with KEGG”超级链接中。

如何才能利用数据库重建一种“生物系统”？KEGG 的方法是将一个生物分出层次，最简单的观点是将其分为原子层次、分子层次和网络（代谢途径）层次。 KEGG 利用了一种数据呈现系统（system for data representation），这种系统根据整体中各组分之间联系数量的多少来建立数据库结构。

数据库目录的组成成分包括分子（蛋白结构、代谢产物）、基因（序列）和基因组。代谢途径图谱通过分子间的相互关系来构建数据库中各组分之间的联系，这些分子间的相互关系包括分子间相互作用（结构）和遗传学相互作用（功能）。两个以上组分之间的相互作用，包括了代谢途径（分子和遗传学）、基因组（线性和环状）、分层（分类、物种）和相邻关系（序列相似性、结构相似性），因此被称为网络。这一点有助于我们对功能重建模型的理解。它是生物信息学应用于蛋白质组学研究中的一项重要内容。

KEGG 以反映生物学实质为目的，介绍了细胞内基因组与信号途径的整和信息。这样，就能使科学家可以将模式生物作为整体，来寻找蛋白质或基因的新的相关信号途径，与单独研究一条信号途径的方法相比有很大的优点。对于一个分子来说，如果它与某个结构已经明确的蛋白质序列同源，这个分子的结构信息就可以预测得到。在构建好的功能模型窗口，选择适当的生物，输入起始点底物和代谢终点的产物来分析，就有可能预测出新的信号途径。

三、大肠杆菌代谢数据库： EcoCyc（E. Coli Metabolic Database）

大肠杆菌是遗传学家、分子生物学家、微生物学家和生化学

家在实验室常用的工具。这是由于大肠杆菌具有繁殖快、易于操作等特点，目前分子生物学实验室中使用的大肠杆菌菌株的遗传特性都已经研究得非常透彻。实际上，实验室中使用的都是致病性非常低的基因工程菌株。大肠杆菌对人类的健康和生理也非常重要，主要是由于它是我们胃肠道中的重要成分。但是，大肠杆菌也是人类致病原。如果它进入了人的血液可以导致致命的感染，许多消化道的疾病也是由于大肠杆菌污染食品造成的。它与另一致病菌—沙门氏杆菌在遗传学上的关系非常密切。因此，了解大肠杆菌的代谢、遗传和致病因子等信息就是一个非常紧要的问题。

EcoCyc 由 Pangea Systems 公司制作（http：//www.pangeasystems.com)，其内容包括了大肠杆菌全基因组序列信息和经典生化代谢途径的整和信息。例如：大肠杆菌中氨基酸合成代谢途径由几种酶参与，它们共同调节基因表达水平。因此敏感的蛋白质组学技术应当能在电泳点中同时看出几种蛋白质的迁移，而不只是一种蛋白质的迁移。

与 KEGG 相似，EcoCyc 应用了化学复合物数据库，其中列出了参与生物学反应的分子及其分子量，并包括了大部分分子的化学结构。

EcoCyc KB 有几种用途。对于研究大肠杆菌和相关微生物的生物学家来说，它是一个电子参考文献来源，科学家在这里可以看到大肠杆菌染色体中基因的排列、单个生物化学反应或完整的生化途径（同时可显示复合物的结构）。用户可以用浏览器从一种酶链接到这种酶所催化的反应，或者编码该酶的基因。网络界面还支持各种查询，例如：可以显示某一生化途径中所有基因的作图位点。

除了作为电子参考文献来源的功能外， EcoCyc 还可以进行与代谢相关的复杂计算，例如：为生物技术设计新的生化途径，研究代谢途径的进化或模仿代谢途径。 EcoCyc KB 还支持建立在计算机平台上的生物化学教育（http：//ecocyc.doubletwist.com/）.

大肠杆菌代谢数据库提供了大量的信息，易于在 PC终端上获得。到 2001 年 6月，在 EcoCyc 的最新版本 5.6 中包含了以下内容：

- 4393 个大肠杆菌基因
- 905 个由这些基因编码的酶
- 2604个大肠杆菌的代谢反应
- 165个大肠杆菌的代谢途径
- 162个转运蛋白
- 629 个转录单位
- 3508 条引文

EcoCyc 制作者曾经这样描述：“大肠杆菌代谢的数字模型可以通过计算机方法来检测和分析”。这意味着利用计算机模拟代谢途径模型可能会代替传统的生化模型实验。因此，未来的电脑可能转变为一个工作平台，成为一个电子实验室。

目前，互联网可被视为一个开放的、没有充分注解和未编辑的数据库。网上有无数的超级链接，找到正确的和完整的信息通常比较困难，有时还是不可能的。一些公司已经认识到有必要对现存的生物信息进行编辑、注解和人为介入工作，以便于生命科学界可以充分利用这些信息，这些公司成立的目的就在于此。对生物信息的注解花费很高，但如果能满足用户的需求，这类产品的销售会很好。这也是高新技术产业一个很好的发展方向。

生命体系是非常复杂的，细胞内各组分在时空上的组织化是细胞功能研究的一个中心问题，但细胞—这个超级分子结构网络还没有搞清楚。这个问题是研究细胞在适当的时间和地点如何组成自身体系的。蛋白质组学研究在某一时刻细胞内蛋白质之间的关系以及与细胞功能的联系。目前，在识别大范围细胞状态下特异表达的蛋白质上有了重要的进步。生物信息学在编辑从微生物到人类的蛋白质表达数据库工作中起到了非常重要的作用。

参考文献：

1. Ogata H, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res, 1999, 27 (1): 29-34
2. Karp, PD et al. EcoCyc: encyclopedia of Escherichia coli genes and metabolism. Nucleic Acids Res, 1999, 27 (1): 55-58
3. Karp PD, et al. EcoCyc: encyclopedia of Escherichia coli genes and metabolism. Nucleic Acids Res, 1997, 25 (1): 43-51
4. Lawrence S, Giles CL. Searching the World Wide Web. Science, 1998, 280 (5360): 98-100
5. Fred W, et al. Biomolecular Mass Spectrometry. Sceince, 1999, 284 (5418): 1289-90
6. Gaskell SJ. Electrospray: principles and practice. J Mass Spectrom, 1997, 32: 677-88
7. Karas M, et al. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem, 60: 2299

第六章　生物信息学在生物学中的其它应用

生物信息学试图从生物数据中提取新的生物学信息和知识，是一门深深植根于实验事实和数据的理论生物学。从目前的发展来看，其应用的范围十分广泛。总的来说，大致包括了以下几个方面：大规模基因组测序中的信息分析；新基因和新 SNPs（单核苷酸多态性）的发现与鉴定；完整的比较基因组研究；大规模基因功能表达谱的分析；生物大分子的结构模拟与药物设计；生物信息的在线服务；生物信息可视化和专家系统。而其长远任务包括非编码区信息结构分析和遗传密码起源与生物进化的研究，读懂人类基因组，发现人类遗传语言的根本规律，从而阐明若干生物学中的重大自然哲学问题，像生命的起源与进化等。如今，生物信息学的应用还见于：汇集与疾病相关的人类基因信息，发展病例样品 DNA 序列信息检测技术，表达载体的选择、引物设计，建立与动植物良种繁育相关的数据库，等等。

关于生物信息学在基因组分析和蛋白质组分析中的应用，本书在相关章节中已有详细的论述。下面就其它领域的应用作一介绍。

第一节　分子结构可视化与计算机模拟

现在，生物学的研究方式有了很大的变化。无论是在大学实验室、私人研究所还是药物及生物技术公司，到处都充满着昂贵的设备和各种专业的工作人员。他们的实验对象可以是一些少量的液体，其中含有生命的基本分子—DNA 和蛋白质。现代分子生物学已成为一门专门研究生命分子的科学。这些分子是肉眼难以

感知的。象孟德尔那样，通过观察描述豌豆的颜色与形状的变化进行研究的时代已经过去了。当代科学家借助新的研究方法，可以检测到影响豌豆颜色和形状的基因位点，并间接地测知其分子结构。人们可以通过对实验数据的数学分析得出基因在染色体上的位置以及蛋白质的分子结构特点。后者即为分子结构可视化的内容。计算机模拟已经进入生物学研究领域，并逐步成为方便的实验辅助工具和训练工具，它对科研方法、工作思维有着深刻的影响。

一、3-D 成像（三维成像）

物质的化学结构，一般可用 X 射线衍射法来探知。高能 X 射线通过某种类型的晶体时，会发生特定的衍射形式。这样，便可根据 X 射线衍射的类型和强度推算出电子在晶体分子内的排布。从推算过程的复杂程度上看，手工计算是望尘莫及的。因此，需借助计算机以阐明蛋白质、核酸、脂类及碳水化合物的分子结构。

客观参数可以反映分子的实际情况。正如水的粘滞力可以反映液态水分子之间氢键动力学和运动颗粒的大小。可通过测量蛋白溶液的粘滞力来测知蛋白质分子的大小和分子量。若采用量子力学论的观点来描述生物分子之间重要的相互作用（如共价和非共价键），将使问题变得十分复杂。但若采用介电常数，这个单一的物理参数来描述水分子的极性，便会大大简化对水中带电分子间相互作用的演算。生物化学中，介电常数常被用来描述蛋白质表面与水分子、其它蛋白质或 DNA 分子表面相互作用引起的能量变化。同样地，膜蛋白也可用反映疏水溶剂介电性的单个参数来研究。这类参数与水溶液的参数有很大不同。对以上参数的选择，实际上是一种行之有效的简化手段，这就是计算机模拟蛋白质分子结构和动力学时将溶剂（如：水）视为一个简单的宏观量的原因。否则，仅仅研究溶剂就需要相当长的时间，并且由于有大量相对独立的参数要考虑进去，对分子结构的模拟很难实现。

在过去，生物大分子的成像是不可能的。1972 年 Levinthal 和

Katz 开创性地建立了计算机线框（wireframe）模型。几年以后，人们利用计算机完成立体分子及其旋转模型的愿望实现了。近年来，已有容量较小的类似 Kinemage 和 Rasmol 的软件可使我们直观地了解蛋白质结构， Rasmol 和 Cn3D 是可免费下载的软件。

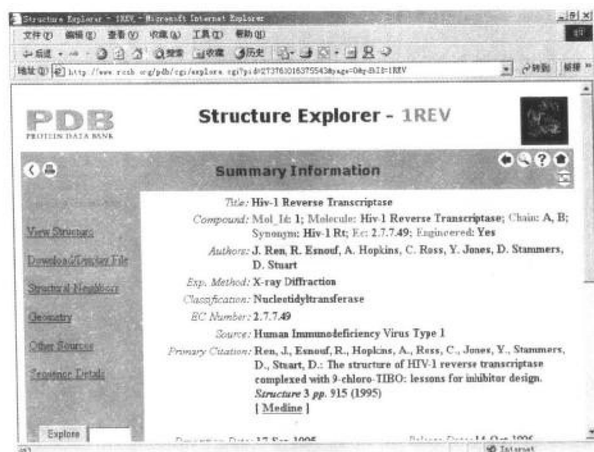


图 6-1 HIV-1 反转录酶的蛋白质数据

图 6-1 显示的是人类免疫缺陷病毒（HIV）的逆转录酶分子。对此分子结构的描述已经达到很高的精度（2.60Å),即描述了除氢原子以外的所有 7715 个原子的相对空间位置。另外，还使用了 X-PLOR3. 1 程序测定了总数为 27108 的反射射线，使结构测定更加精细。应用该酶的序列号 1REV 可以查询 Brookhaven 蛋白数据库（PDB；http://www.rcsb.org），查询结果显示其提交的日期是 1995 年 9 月 17 日，还有蛋白名称(HIV 逆转录酶),以及作者姓名(J. Ren 等），同时还可以显示软件呈现的结构图，见图 6-2.

Compound 和 Classification 列出的信息使人们知道该蛋白源自人类免疫缺陷病毒 I 型的反转录酶。该反转录酶是由基因工程重组表达的，其国际酶学委员会编号为： EC2.7.7.49。为了获得

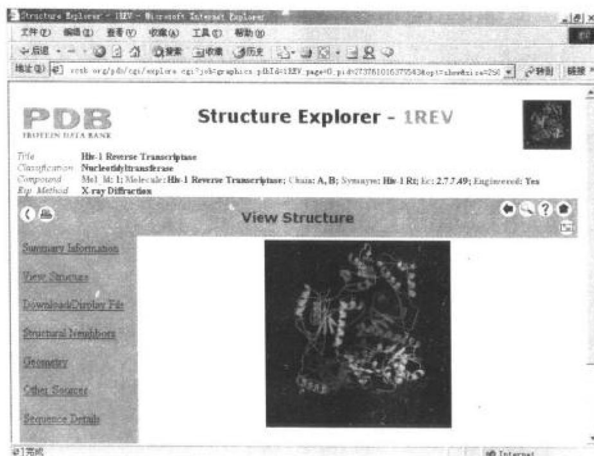


图 6-2 HIV-1 反转录酶的结构图（网页中为彩色）

最大量的结晶，可以将该酶表达于细菌。一般所选的表达体系为大肠杆菌，重组 DNA 中包含人类免疫缺陷病毒 I 型反转录酶基因，使其编码的蛋白过表达，有利于蛋白纯化。当然，也需明确蛋白的结晶参数、方法，外加配体及解析度（2.60 Å ）。为了对目前该蛋白的结构有个全面的认识，科学家们可通过一些网络链接去评估该蛋白的单层扩展（spreadsheet）模型与 3-D 及 2-D 图像之间的相关性。点击“Other resources ”的链接，可获得 EBI MSD Macromolecular file sever、SCOP、MMDB 等数据库中的序列信息以及其它相关信息的链接（可得到 MEDLINE 中的文章）。

使用 MDL 的 Chemscape Chime 浏览器可轻松地以不同模式观察分子的结构。另外， PDB 结构图形文件还可下载，并可随时使用数据 Chime 软件的 Rasmol 浏览器浏览。

附：虚拟医生（Virtual Doctor）及虚拟人体

在先进的临床医学诊治手段中，计算机和机器人常是重要的组成部分。这些诊治手段高效省时其日益提高的精确性正驱动着它们在当今和未来的

医学中的快速发展。所谓虚拟医生，一般是指虚拟的或远程的外科手术及其它治疗。模拟仿真软件在一些医院里被应用于外科手术演练和培训。虚拟手术虽有助于提高某些手术的成功率，但不能预测真实手术中可能出现的一些问题。而正是对处理意外情况的不断训练，会大大提高外科医生的操作水平以及在真实手术时的治疗效果。因此，仿真软件的拟真度是至关重要的。要使手术区的结构得以真实地展现，对各种解剖部位三维结构特征的再现必须十分精确。虚拟外科手术的难度必须能够测知，以便于量化术前手术操作水平，便于培训时的自我控制。

数百年来，人体解剖结构成像一直是医学界的一大挑战。1489 年 Leonardo da Vinci 详尽的解剖图谱在当时对医学教育、诊断和治疗都有巨大的影响。但 Leonardo 的图谱却不能显示存活着的人体结构。直到 1895 年 Conrad Roentgen 发现 X 线，人们的这一梦想才有了可能。 X 线检查是第一种用于临床的医学活体成像技术。该技术的局限在于，它仅能呈现横切面的二维结构，而这样收集到的组织器官的信息量很少。尽管如此，直到今天， X 线仍是医学实践中很有用的工具。上个世纪 70 年代发展起来的计算机断层扫描成像（CT）和 80 年代发展起来的核磁共振成像（MRI）,大大地提高了活体解剖结构断层的成像效果。由这些技术获得的二维图像有助于临床诊断和解剖学教学质量的提高。

二维图像的缺点是缺乏立体上的信息，它不能显示物体在三维空间上的特征关系。三维图像可表现组织结构之间相互关系的信息（如：组织-组织、器官-器官）。在二维图像中，获取的信息可建立在两个参数的坐标系统上。要使二维的图片连接成三维图像，必须有来自第三维度（Z 轴）的信息。现在，有几种三维重建技术被用来提高图像的质量和真实性，它们使展示的图片对观察者更具吸引力。“灰度梯度显影法”（gray level gradient shading）、“Generalized Voxel Model”和人体模型可视化的重建是近十年来最为成功的几种技术。利用灰度梯度法，计算机依原始的数据断层图像计算产生动力范围的光滑面标准。而“Generalized Voxel Model”则可进行图像数据信息的进一步分析处理，该技术的应用之一是能同时显像各个器官并可选择性截取三维图像。最近的研究集中于构建虚拟的“可视人体”。在这一项技术中，彩色断面成像法提高了物体三维图像的质量和真实性。计算机已经并将继续推进医学的发展。

二、虚拟细胞与预测生物学

目前，生物信息学最活跃的前沿是基因组信息学，它正在成为发现基因、破译基因组密码并推动实验科学的强有力工具。基因组信息学的首要任务之一就是发现新的基因和功能。由于测序技术的进步，各国政府和公司资助的测序中心首先对重要病原体和工程菌的基因组进行测序。至今已有多种重要微生物的全基因组的测序工作已完成。

这些细菌基因组的全序列立即成为研究新的更精确有效的诊断、治疗手段和新药物的基础。以前，分子生物学家一直习惯于分离和分析一个个的基因。随着生物学知识的积累和计算机技术的发展，研究细胞全基因组在生理和各种病理过程中表达的动态变化已经成为可能。这是实验生物学进步的必然，也将为理论生物学成为整个生命科学的先锋带来了重大的机遇。在日本的 Keio 大学有一个名叫 Masaru Tomita 的生物信息学教授领导的研究小组，正在做一个有着划时代意义的软件： E-CELL ,我们称之为虚拟细胞（见图 6-3）。这是一种生物学计算机模拟软件，在计算

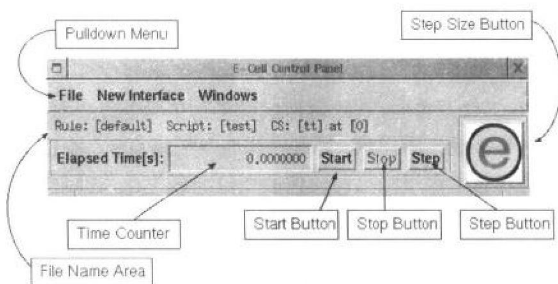


图 6-3　E-Cell　软件的控制面板

机环境中构造一个虚拟的电子细胞。它不仅仅包括一些单一的细胞事件和过程，而是从整体的角度为细胞描绘一幅全图。电子细胞将把每个时刻特定位置上特定物质的变化，通过画面和数字告诉你。研究者可以仅仅用鼠标去轻轻点击，就能实现在分子生物学实验室花费大量时间和金钱进行的基因敲除、转基因或基因修

饰等操作，自由地将感兴趣的细胞暴露在某一种生存环境下，而无需考虑细菌的污染、RNA 的降解或放射性损害。研究者所需做的就是输入初始值，然后就是在计算机屏幕前等待 E-CELL 的模拟结果。毫无疑问，这种方法将提供一个非常简捷经济的手段，来进行药物筛选和基因功能研究。更重要的是，我们能实时的看到某个因素和环节对细胞整体行为及生命活动的影响。目前这个程序可在 UNIX 或 Linux 操作系统下运行。Tomita 的研究小组已经开发了 E-CELL 的 Windows 版本，可以通过 http://bioinformatics.org/project/网址的相关链接获得。

Tomita 的小组已用 E-CELL 的早期版本建构了一个“假想的细胞”，拥有大部分来自解脲支原体（最简单的细胞和最简单的基因组）的 127 个基因。这个虚拟的细胞就在计算机环境下“生活”着，从虚拟的培养基中吸取着葡萄糖等养分，合成各种各样的维系细胞生存的酶和蛋白质，排出乳酸等代谢废物。它的重复性很好，绝没有人为的误差；更重要的是，它在给我们一种崭新的探索环境，我们能从已知里寻找未知的联系，检验我们的思想。

除了基因组信息学外，生物信息学在测定蛋白质结构方面，也有着突出贡献。利用分子模拟技术结合计算机图形技术可以更形象、更直观地研究蛋白质等生物大分子的结构。当前的分子模拟技术主要借助于先进的计算机图形工作站，通过友好的图形环境，使用者可利用鼠标极为方便地建立多肽、蛋白分子的初始模型。同时，也可以对已知的生物大分子的三维结构进行显示，并对这些结构进行灵活方便的平移、旋转、放大及缩小等操作。分子模型的建立为下一步进行的分子模拟以及了解结构与功能的关系打下了基础。

生物学和医学最新的研究进展给这些领域增添了许多预测性的因素。利用生物信息学预测工具在蛋白质结构预测和药物设计工作中作出了很大的贡献。这就出现了一门新的学科一预测生物学。这门学科利用以往的一些研究结果（如：已知的蛋白质一级

结构和三级结构关系）总结出其中内在的规律，利用这些规律去预测一些可能的结果（如：蛋白质的三级结构等），并可以利用这些结果去指导将来的实验研究。

近年来，国际上一些研究组发展了一些从蛋白质的一级结构直接预测蛋白质空间结构的新方法。这些方法的基本思想是将基于生物学知识的方法与计算化学以及统计物理学的方法相结合，采用简化的蛋白质模型和根据已知结构的蛋白质所导出的平均势场，从理论上计算蛋白质的空间结构。这些方法不仅可以从蛋白质的一级结构直接预测蛋白质的三维结构，而且可以在计算机上模拟蛋白质分子折叠的全过程。目前，还有一些新方法，如遗传算法、模拟退火、多维统计、模糊集合论等方法在蛋白质结构预测中的应用也正在研究之中。通过对一些简单蛋白质分子的模拟研究，这些新方法已经显示出很强的生命力。许多权威人士推测，随着这些新方法的进一步改进和完善，在今后 10 年内，蛋白质折叠这一分子生物学中的难题将有望得到解决。

另外，生物信息学在药物设计方面也有着广阔的发展前景。传统的药物研制主要是从大量的天然产物，如动物、植物、微生物和合成有机、无机化合物以及矿物中进行筛选。得到一个可供临床使用的药物要耗费大量的时间与金钱。近年来由于生物信息学的发展，相当数量的蛋白质以及一些核酸、糖类三维结构已被人们精确测定，使得基于蛋白质和核酸结构的药物设计成为可能。比如近年开发的用于药物分子初期设计的 LUDI 软件，人们只要将所感兴趣的受体及可能的药物分子结构输入， LUDI 就可以计算该药物分子对受体抑制活性的相对值。这种评估方法可对作用于某一受体并具有不同抑制活性的大批药物分子进行快速筛选。

第二节　神经生物信息学的研究

人类基因组计划对人类健康、疾病诊断、药物开发、生态平

衡和生物学研究有着不可估量的贡献。许多科学家认为，在人类基因组计划之后应该是人类蛋白质组计划和人类脑计划。

人类脑计划的核心内容是神经生物信息学。神经生物信息学是脑科学和信息学这两大学科相结合的新兴的边缘学科。其目标是利用现代化信息工具，使神经科学家和信息学家能够将脑的结构和功能研究结果联系起来，建立神经生物信息学数据库和有关神经系统所有数据的管理系统，将不同层次的与脑研究相关的数据进行检索、比较、分析、整合、建模和仿真，绘制出脑功能、结构和神经网络图谱，从而解决目前神经科学所面临的大量数据问题，从基因到行为各个水平加深人类对大脑的理解，达到“认识脑、保护脑和创造脑”的目标。

人脑的复杂性远远超出了我们目前的认识能力。传统的神经生物学等实验室研究对于解决人脑对复杂信息的获取、处理与加工及高级认知功能的机制，犹如只见树木不见森林。神经生物信息学工具和数据库的应用，使得我们可能从有限的实验数据中找出神经信息获取、处理和整合的规律和法则，提出在各种刺激条件下，脑内信息加工的数学模型的实验假设和用计算机模拟脑内神经信息网络。人类脑计划的发展与神经信息学紧密相连。

一、人类脑计划和人脑图谱

由于研究行为、意识、记忆、睡眠紊乱、感觉或疼痛（包括四肢痛）等的需要，大脑的研究受到了科学家们的高度重视。最近，计算机辅助脑扫描技术和人类基因组计划的成功，再次促使人们去尝试描绘大脑的解剖功能图谱。有了这种定位图谱，人们终究会有一天，可在单个感觉神经元水平把神经元的活动和其功能同时定位。新兴的学科——神经生物信息学将人类基因组计划的最新研究进展应用到神经科学的研究当中，即 DNA 序列与脑功能数据资料间的结合。这种结合为研究人类的行为方式的机制提供了解释。

1993 年 4 月 2 日，美国 NIH 正式宣布人类脑研究计划

(Human Brain Project ）启动。人类脑研究计划作为一个多学科的长期性研究，开创性地支持了先进技术手段的研究和发展，为神经科学家和行为科学家们打开了信息高速公路—互联网之门（人类脑研究计划，http://www-HBP.scripps.edu)。美国 20 多家著名的大学和研究所参加了这个研究计划。 50 多位神经信息学的课题负责人得到该项目的基金资助。他们充分利用神经科学和信息科学的优势条件进行研究，相互间建立合作关系，利用电子网络互通信息，运用数据库进行资源共享。

1996 年在巴黎的政府间实体—经济合作发展组织(OECD)的科学论坛上，批准建立以美国为领头国家的神经信息学工作组，参与国包括：美国、英国、德国、法国、瑞典、挪威、瑞士、澳大利亚、日本等 19 个国家，欧洲委员会也作为正式成员参加。其目的是组织和协调全世界神经科学和信息学家共同研究脑、开发脑、保护脑和创造脑。根据规定，成员国之间可利用电子网络寻求研究协作伙伴，进行数据交换和科研协作，可以免费使用通用神经信息学数据库和信息工具，承担科研任务，共享科研成果和脑研究资源。

人类脑研究计划包括三个子计划：

- 链接计划：多模型成像和神经元链接分析（the Multi-Model Imaging and Analysis of Connectivity）
- 脑图谱计划：活体脑发育图谱(the In Vivo Atlases of Brain Development）
- 算法计划：三维分析和可视化的算法（Goal-Based Algorithms for 3-D Analysis and Visualization）

许多不同技术需应用于该计划：如化学、动物模型、计算机技术、网络工作、网页设计、功能核磁共振成像等技术。为了更清楚地了解人脑的复杂性，脑研究计划的最终目标是将脑研究中的神经科学内容与信息学内容“编织成一个整体”。以下为加州理工学院的观点：

人类脑研究计划的进步，是在各自独立的研究计划和研究中心之间互动的作用下取得的。许多研究人员的专长和兴趣跨越了不同的研究领域，从而促进了脑研究的进展。信息学部分的工作为神经学部分提供了收集、分析和观察信息新颖而更有效的的方法。信息学家们也有必要向神经科学家们学习这些数据是如何收集的，以及什么样的信息是重要的、需要引起关注的（http://www.gg.caltech.edu/hbp/）。

现在，有许多不同学科的专家——从分子生物学、电生理学家到认知学家、哲学家——都加入了神经生物学的研究。其中有些科学家从事与意识有关的认知学问题的研究，他们相信人们最终可以在分子水平上撩开“意识-机体”(mind-body）问题的神秘面纱。脑研究是由各类专家承担的，其中每位专家都有其独特的研究手段、研究技术和研究方法。从生物物理学到心理学的诸多学科之间没有一种“通用语言”，但许多科学家依然相信这种通用语言存在。人类意识的两个主要方面：情感和思维，从生物学角度看具有“难以琢磨”的特点。的确，某些化学物质可以影响人们的意识，似乎意识是存在于中枢神经系统之中的。然而，人们对于脑内的化学反应在时空上是如何产生意识的，却所知甚少。生命科学中许多不同领域间的信息鸿沟阻碍着神经生物学的发展。

人不能看见，或不能感知自身体验以外的物质世界。为了克服这一天生缺陷，我们必须创造出人类可以感知的“影像”——将那些不能感知的事物转换成为可视世界。伪色成像技术（False color imaging）就是这样一种功能强大的工具，可以将物理参数单位转换为从红到蓝的颜色代码。例如，将温度梯度转换为从红到兰的颜色代码。这样一来，人们不必去阅读和比较数字，大脑就会自动地辨别出个体之间在空间分布上的差别。这样，我们便可“看见”温度梯度（可视化红外线成像），或脑中的分子氧耗量以定位脑中高代谢的活动点。伪色成像是一项令人着迷的技术，它

可使人们把抽象的数学方程进行转化，把各种数值转化为眼睛的视觉感受。计算机推进了这种转化在生命科学研究中的应用。

Ramony Cajal(1852-1934)是一位西班牙的神经解剖学家。早在1906年，因其出色的工作获得诺贝尔医学奖。他细致而又精确地记录了神经系统的结构图，这一工作充分体现了19世纪科学研究的精致与严谨。Camillo Golgi是1906年诺贝尔医学奖的另一位获得者。他发明了一种新的神经细胞染色法，可以染出单个细胞或细胞群，并显示出神经元之间的连接。在这一技术的基础上，Cajal对神经元作了艺术般的“描绘”，因而成为当时脑科学研究的代表人物。Cajal的图谱显示：脊椎动物大脑是由数以亿计的单个细胞和神经元组成的，而不是由细小动脉连接成的网络。今天，人们采用新的染色技术、成像技术与计算机辅助相结合的方法正在构建人脑解剖图谱（参阅哈佛医学院全脑图， http://www.med.harvard.edu/AANLIB/home.html)。

功能性脑图谱的绘制（Functional Brain Mapping）采用无创技术如SPECT/PET、fMRI、EEG、MEG、视觉成像和神经解剖学方法。这些工具用来绘制脑的横断面图谱。横断面图再一起重组成脑的三维图像。人脑整合图谱为解剖和功能图谱添加上了时空维度。

在空间上fMRI对脑结构功能在毫米水平的分辨率和在时间上EEG、MEG毫秒级的分辨率，需要用计算机手段将二者联系起来，以形成脑活动的电影图像。哈佛大学医学院（http://www.med.havard.edu/AANLIB/home.html）的全脑图是向公众开放的，它提供正常脑图和脑血管疾病（如中风）、增生性疾病（如肿瘤）、变性疾病（如阿尔茨海默病，亲壬顿病）及炎症性或感染性疾病（如多发性硬化、AIDS相关性痴呆、Creutzfeld-Jakob综合症（人疯牛病）、疱疹脑炎）的脑图。空间上毫米级的分辨率远远不能达到在分子层次上对脑功能研究的要求。尽管神经元的轴突可长达数毫米至数米不等，但其胞体却仅有数微米大小，神经元

的化学突触则更小。这样，神经解剖图谱（诸如全脑图谱）无法达到在分子水平上的细节描绘。这些细节来自于生物化学，生理学、药理学和分子生物学的研究，如离子通道、受体分布（蛋白组学）、mRNA 水平分布（基因组学），等等。尽管目前全脑图谱的信息容量和分辨率水平能够精确地指出运动神经中枢，但还不能提供单个神经元或神经元群的电活动方式和其对神经递质的选择性。换句话说，结构细节还未与脑的功能状态和当时的意识活动联系起来。

大脑是生物体内结构和功能最复杂的组织，也是极为精巧和完善的信息处理系统，掌管着人类每天的语言、思维、感觉、情绪、运动等高级活动。人类脑计划可以使人们对这个高度发达的处理系统有一个较为深入的认识。

二、神经变性性疾病的分子机制

动作电位是在细胞膜局部表面上维持数微秒的微小跨膜电压变化。分子神经生物学关注与动作电位相关的蛋白复合体，即所谓的离子通道的结构功能关系。离子通道是离子穿过细胞膜的通路。细胞膜是电绝缘的，在没有通道存在时阻滞带有正电或负电的离子通过。离子流可用电流来测量，利用高敏感的电流放大器，便可检测出千分之一秒内几千个离子的运动，因而也可揭示整个细胞膜或部分膜片电活动的生物信息。神经元信息传递的单元一动作电位，是至少三种不同离子通道共同活动的结果。它意味着对每一种离子，细胞上必需存在与之相应的一种蛋白允许其跨膜流动。众所周知，离子通道具有离子选择性。

在神经生物学中，单个离子通道的知识绝大部分来自对单细胞电流的生物化学和动力学分析。离子通道的共同活动形成了神经细胞的宏观行为，这要求对膜离子通道活动要有正确的数据表述。

许多这类通道蛋白与神经变性性疾病及大脑对药物和毒物的易感性有关。虽然自 20 世纪 40 年代以来，人们已对离子通道进

行了功能和结构上的详尽的研究，但 90 年代采用分子生物学技术对新基因及通道蛋白的基因结构的研究，极大地推进了人们对诸如阿尔茨海默氏病、帕金森氏病及亨廷顿病的认识。人类基因组计划将会加深我们对这些疾病及其它遗传性疾病的理解。寻找如何治疗这些疾病的研究为生物信息学在不同的研究层次上提供了用武之地。

例如，QT 间期延长综合征是人类心肌节律控制失调的一种表现，可出现突发性意识丧失和心源性猝死，常见于受各种应激（如重体力活动）的青少年。这种由心肌细胞钾离子通道结构异常导致的疾病往往表现为快速心律失常。这里起主要作用的是选择性钾离子通道，它排斥其它任何具有生物学意义的离子，如钠、氯、钙、镁等通过。通常细胞内钾离子浓度高出细胞外十倍，一旦钾离子通道开放，钾离子从胞内流向胞外，使胞外带正电。这种钾离子外流将一直持续到细胞内外钾离子浓度相同为止。在正常体细胞中，这种现象不会出现。

体细胞不能达到这种平衡的机制有两种解释：首先，细胞能把钾离子泵入胞内，同时，通过同一转运蛋白一钠钾泵一把钠离子泵出胞外，造成钠离子膜内外的不对称分布，产生一种对抗钾离子流的电场力。其次，通常钾离子通道在短暂的开放后，受到调控而长时间关闭（失活）。钾离子通道的开放时间及离子流量因细胞不同而异，说明不同的细胞上存在着不同的钾通道。这可能是因为细胞基因组上存在着不同但又功能相关的钾通道基因，这些基因的差异性表达造成了组织和细胞钾通道分布和活动的特异性。单纯钾通道活动的结果是恢复神经元的静息电位（复极化），在 QT 间期延长综合征中，一种钾通道基因发生突变，就会引起复极显著减慢，导致心肌节律失常。

目前，通过逐个克隆的办法发现了许多钾通道的基因。通过序列分析和比较，我们更深入地认识了这些基因的作用及生物体如何利用它们所编码的蛋白质。在未来生物学的发展中， DNA 和

氨基酸的序列分析与功能信息相结合是极为重要的。将基因多样性、细胞特异表达和功能相结合，可建立基因功能图谱数据库。

目前已克隆成功并完成测序的钾离子通道基因已超过 50 种，这一数目包括了从人到细菌多种生物的相关基因。随着测序的进展，我们能对一些功能信息已知的序列差异做出有趣的解释。再者，对序列、结构、功能间关系的理解有助于我们对这些通道的正常生理和相关疾病的认识。

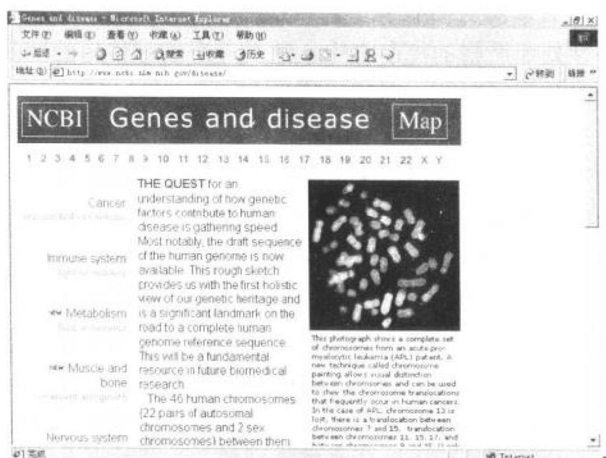


图 6-4 Genes and Diseases：NCBI　关于遗传性疾病基因的数据库

基因测序和作图在确定疾病相关基因方面的进展在 NCBI 站点有详细的介绍（图 6-4）。NCBI　（美国国家生物技术信息中心）提供了基因与疾病的信息总汇。八种遗传性疾病与离子通道、泵和转运蛋白缺陷有关。包括囊性纤维化（与氯离子通道有关），形态不良性不典型增生（diastrophic dysplasia，与转硫蛋白有关），QT 间期延长综合征（与钾离子通道蛋白有关）， Menker's 综合征（与铜离子转运有关），Pendre 综合征（与胸腺特异的转硫蛋白有关）、多囊肾（与细胞间连接蛋白及膜蛋白组织有关）， Wilson's disease　（与铜离子转运蛋白及 ATP 酶有关），Zellweger　综合症

（与 PXR1 蛋白有关，为过氧化物酶蛋白转运受体）。

相同物种，甚至不同物种间离子通道研究的生理学、药理学、分子遗传学、种系生物学的分析与比较也是分子生物学未来的发展方向之一。

第三节　生物信息学在肿瘤学研究中的应用

肿瘤通常发生在细胞分裂失去控制的情况下。正常细胞分裂的时序性处于严格的控制之中，这是一个信号传递的网络。它决定着细胞分裂的时机、时间间隔以及如何修复分裂中的产生的错误。在这个网络中，由于环境因素（如：吸烟）或遗传倾向引起的单个或多个关卡基因的突变都有可能导致肿瘤的发生。大部分肿瘤是由几种促癌因素一起作用的结果，而单个的促癌因素一般不会引起肿瘤发生。肿瘤发生的机制可以归纳为以下三点： 1. DNA 修复途径的损伤，2. 正常基因转化为肿瘤基因。 3. 肿瘤抑制基因失活。总之，肿瘤的发生与基因的变异是紧密联系的。

在研究肿瘤发生机制的过程中，通常使用比较正常细胞和肿瘤细胞之间的差异的各种方法，例如形态学的差异、基因表达谱的差异以及蛋白质表达谱的差异，等等。而后两者的比较研究常常建立在对大量生物数据排列分析的基础上，数据的繁复促使人们寻求计算机来辅助完成。生物信息学在这方面提供了可靠的分析工具。下面试举一例(图 6-5）来说明生物信息学在肿瘤学研究中的重要性和方法。

例如：研究某种白血病的发病机制，可以选择该病的罹患者作为研究对象。采取肿瘤细胞和正常对照细胞标本，提取两者的 mRNA，通过反转录和差异显示的方法分离出两者之间表达差异的基因，这些差异的基因经过测序后得到其序列。序列数据就可以应用大量的生物信息学工具来分析。对于上述实验中得出的核苷酸序列，首先要明确它是一个包含了全编码区的基因，还是一

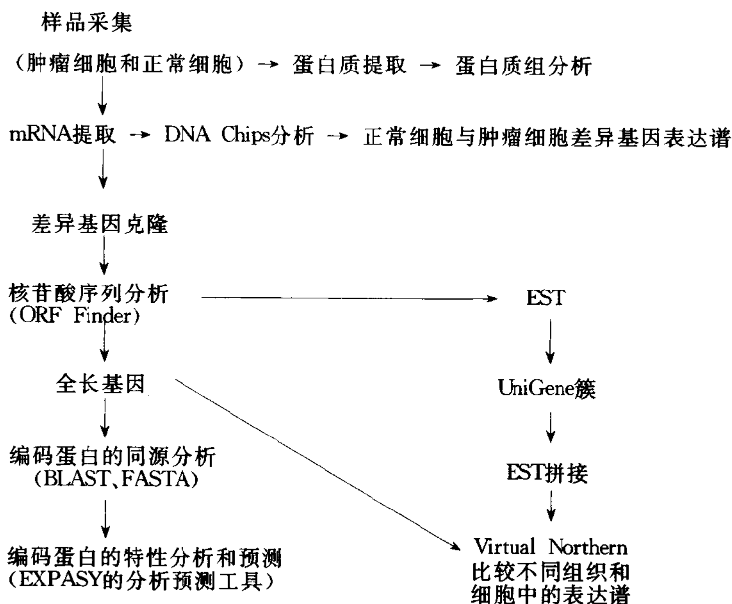


图 6-5 生物信息学在肿瘤学研究中的应用方法示例

个基因的部分序列，这就需要使用类似 ORF finder 的软件来查找序列中是否存在 ORF 以及 ORF 的长度。另外，还要识别 ORF 5'端和 3'端序列中的特征，如：接近起始密码子的序列特征是否符合真核生物基因的规律，3'端序列中是否存在 PolyA 加尾信号以及序列中包含何种核酸motif 等。这些特征都是识别全编码区的重要标志。许多公司的商业化软件包都提供了这些功能，如：OMIGA 和 DNASTAR 软件包。这些软件包中还有将核酸序列翻译为蛋白质序列的功能，互联网上也有类似的工具，如：EBI 提供的翻译工具 Translation Machine (http://www2.ebi.ac.uk/translate/)。

如果经分析后发现该序列是一个基因的部分序列，那么这个序列就是一个表达序列标签，它代表一个基因。这个序列可以通

过 NCBI 的 BLAST 工具查询 EST 数据库来寻找同源的 EST 序列。利用同源 EST 条目的序列号可以查询 UniGene 数据库，这样可以得到该序列对应的 UniGene 簇（EST 数据库中同源的序列总结为 UniGene 簇，每个 UniGene 簇代表一个基因）。EST 序列还可以用于在 EST 数据库中进行序列拼接，以得到更长的序列，甚至可以得到包含全编码区的序列，这个过程称为 "in silico cloning"。英国 Human Genome Mapping Project Resource Centre 的生物信息学服务提供了 ESTBLAST 软件（http://www.hgmp.mrc.ac.uk/Registered/Webpp/estblast/），这个软件可以将输入序列与 EST 数据库比较得到同源序列，并通过排列将这些同源序列拼接为重叠群得到更长的序列。

如果分析后发现该序列是一个包含全编码区的序列，那么该序列中包含的开放读框对以后的生物信息学分析和预测非常重要。开放读框的核苷酸序列可以翻译为氨基酸序列，氨基酸序列可以通过 BLAST 或 FASTA 同源搜索工具在蛋白质数据库中查找同源的蛋白质。这样，就可以明确得到的氨基酸序列是否是一个新的序列，另外通过了解同源蛋白的功能还可以预测该蛋白的功能。

经翻译得到的氨基酸序列可以用来分析蛋白质的许多特性，EXPASY 网站（http://www.expasy.ch/tools/）中提供了大量的蛋白质分析软件的链接，其中包括蛋白质一般特性分析、DNA 翻译工具、序列相似性搜索、蛋白质特征（Prosite）分析、转录后修饰预测、一级结构分析、二级结构预测、三级结构分析、跨膜区预测和序列排列等十大类分析工具。我们可以通过这些工具得到蛋白质的大量信息，例如：分析蛋白质中可能与 MHC I 类分子结合的肽段可以指导以后的抗原肽实验；分析蛋白质中是否有信号肽可以预测蛋白质是否是分泌蛋白；分析蛋白质中含有的蛋白酶切割位点可以预测蛋白质的半衰期；多序列排列可以明确蛋白质之间的进化关系，等等。每一类中都包括多种分析工具，在同

一类分析中可以使用不同的工具，分析结果可以相互比较得出最可靠的信息。

由于人类基因组计划已经完成了大部分的测序工作，所以分析基因的染色体定位和基因组结构就可以通过同源比较的方法来实现。我们可以将实验中得到的基因作为探针进行 Northern Blot 杂交。如果杂交结果中的转录子长度与实验中克隆的基因长度非常接近，这个基因的序列就接近全长的 mRNA 序列，这样的 mRNA 序列就适合用来通过同源搜索进行染色体定位和基因组结构分析。方法非常简单，将 mRNA 序列与基因组数据库进行 BLAST 同源比较就可以完成。

某个基因在一些组织和细胞中的表达情况也可以通过生物信息学工具得到。NCI 的 CGAP 计划中提供了一个称为 "Virtual Northern" 的工具，这个工具可以分析基因在 NCBI 构建的正常和肿瘤组织或细胞系 SAGE 文库中出现的频率。这些不同的频率体现了该基因在不同组织或细胞系中表达水平的高低，这可以反映出基因在肿瘤发生中的作用。具体操作方法是：将基因序列粘贴到 Virtual Northern 页面（http://www.ncbi.nlm.nih.gov/SAGE/sagevn.cgi）的窗口中，然后选择限制性内切酶（NlaIII 或 Sau3A），点击 "submit" 链接后可以得到查询结果。查询结果中显示的标签是 SAGE 标签，它应当符合输入序列的标签，输入序列的 SAGE 标签是由 Virtual Northern 软件自动计算选择的。最终结果是 SAGE 标签在不同的 SAGE 文库中出现的频率，它代表了含有该标签的基因出现的频率，可以反映该基因在不同组织或细胞系中的表达活性。

比较正常细胞和肿瘤细胞之间的基因表达谱差异还可以使用基因芯片的方法。基因芯片（gene chip）也叫 DNA 芯片、DNA 微阵列（DNA microarray）、寡核苷酸阵列（oligonucleotide array），是指采用原位合成（in situ synthesis）或显微点样手段，将数以万计的 DNA 探针固化于支持物表面上，产生二维 DNA 探

针阵列，然后与标记的样品进行杂交，通过检测杂交信号来实现对生物样品快速、平行、高效地检测或医学诊断。由于常用硅芯片作为固相支持物，且在制备过程运用了计算机芯片的制备技术，所以称之为基因芯片技术。应用这一技术要先分别提取肿瘤细胞和正常细胞的 mRNA ,标记荧光制备探针,再与 DNA 芯片杂交，杂交信号经专门的检测仪器阅读后经计算机分析得出结果。这种方法可以在一次实验中分析数千至数万种基因在肿瘤细胞和正常细胞中表达的差异。Stanford 大学提供了专业的 DNA 芯片数据库 Stanford Microarray Database（SGD，http：//genome-www5.stanford.edu/MicroArray/SMD/）。目前，DNA 芯片在肿瘤研究中的应用非常广泛，并且产生了许多重要的科研成果。一个很好的例子是美国 Dana-Farber 癌症中心的研究人员利用基因芯片鉴定出一种新型白血病，命名为混合谱系白血病（Mixed Lineage Leukemia，MLL）。这种白血病的特征是最初对化疗有一定的反应，但一旦复发却可致命，预后很差。以前，临床医生将这种白血病归类到急性淋巴细胞性白血病（ALL）之中。研究人员应用基因芯片技术证实混合谱系白血病的基因表达谱与普通急性淋巴细胞性白血病有明显的不同。与 ALL 患者相比，MLL 患者细胞中有将近 1000 个“沉默基因”或处于失活状态的基因，同时还有 200 个基因处于过度激活状态。这些研究结果对于这种特殊白血病的诊断和治疗都非常重要。

研究肿瘤发生机制还可以使用蛋白质组学的方法，这种方法可以用来比较肿瘤细胞和正常细胞之间蛋白质表达谱的差异。EXPASY 的二维凝胶电泳数据库提供了大量的肿瘤细胞和正常细胞蛋白质表达谱信息，可以用于这方面的研究，有关内容在第五章第一节中有详细介绍。

在科学技术的发展中，各种学科之间相互借鉴产生的交叉学科、边缘学科表现出蓬勃的生机。就生物信息学自身的研究而言，

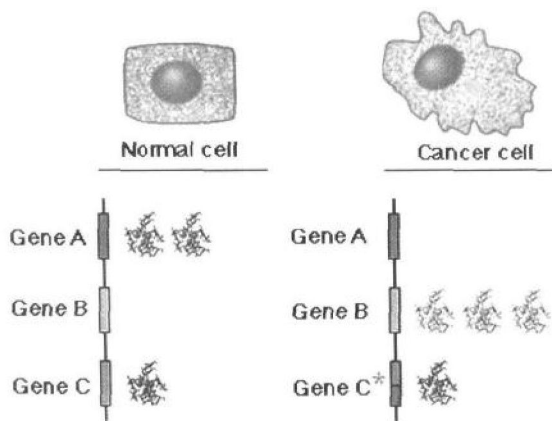


图 6-6　正常细胞和肿瘤细胞基因表达谱的差异，在图中，与正常细胞相比，基因 A 在肿瘤细胞中表达下调，而基因 B 在肿瘤细胞中表达上调，而基因 C 在肿瘤细胞中发生了突变（摘自 http://cgap.nci.nih.gov/Info/concept4）.

随着人类基因组计划等大规模测序计划的进行，随着更多的数学、统计学、计算科学、数据库技术被借鉴进来，随着更多、更详细、更准确的生物学知识整合进来，生物信息学将为揭示生物世界的奥秘，为利用生物资源提供新的手段做出应有的贡献；另一方面，生物信息学作为传统分子生物学的辅助工具的作用不容忽视，生物信息学工作者应该深入到实验室中去，了解各种实验操作的原理、方法，从中找到开发新数据库、软件的线索，帮助分子生物学家更高效地完成实验。这两方面相辅相成，密不可分。

总之，生物信息学作为一门新兴学科有着美好的前景，也面临着巨大的挑战，如何使她发挥更大的作用，是我们面前的一个诱人的课题。

参考文献：

1. Shepherd GM, et al. The human Brain Project:

neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscinece data. Trends Neurosci. 1998. 21 (11): 460-468

2. Yang WP, et al. KvLQT1, a voltage-gated potassium channel responsible for human cardiac arrhythmias. Proc Natl Acad Sci USA, 1997, 94 (8): 4017-4021
3. Badger J, et al. New features and enhancements in the X-PLOR compite program. Proteins, 1999, 35 (1): 25-33
4. Sussman JL, et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr, 1998, 51 (1): 1078-1084
5. Benson DA, et al. GenBank. Nucleic Acid Res, 1999, 27 (1): 12-17
6. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acid Res, 1999, 27 (1): 49-54
7. Richardson DC, Richardson JS. The kinemage: a tool for scientific communication. Protein Sci, 1992, 1 (1): 3-9
8. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci, 1995, 20 (9): 374
9. Cao QL, et al. Enhanced comprehension of dynamic cardiovascular anatomy by three-dimensional echocardiography with the use of mixed shading techniques. Echocardiography, 1994, 11 (6): 627-633
10. Slavin KV. The visible human project. Surg Neurol, 1997, 48 (6): 638-639
11. **阎隆飞、孙之荣，蛋白质分子结构，清华大学出版社，** 1999
12. **吕秋军、高月，受体药物筛选研究进展，中国药学杂志，**1999,

1: 6-8
13. **张亮仁，以结构为基础的药物设计与分子模拟，药物学研究与展望，科学出版社，** 1999
14. **吴旻，生物信息学的发展，中国科学院院刊，** 1998，3：183-185
15. Tatusov RL，et al. A genomic perspective on protein families. Science，1997，278：631-637
16. Koonin SE. An independent perspective on the Human Genome Project. Science，1998，279：36-37
17. Armstrong SA，et al. MLL translocations specify a distinct gere expression profile that distinguishes a unique leukemia. Nat Genet，2002，30 (1)：41-7

附录一　分子生物学数据库一览表

数据库名称和分类	互联网网址	内容介绍
主要序列储存数据库		
DNA Data Bank of Japan (DDBJ)	http://www.ddbj.nig.ac.jp/	包括所有核酸蛋白序列,国际核酸序列数据库协作组成员
EMBL Nucleotide Sequence Database	http://www.ebi.ac.uk/embl.html	包括所有核酸蛋白序列国际核酸序列数据库协作组成员
GenBank	http://www.ncbi.nlm.nih.gov/	包括所有核酸蛋白序列国际核酸序列数据库协作组成员
Genome Sequence Database (GSDB)	http://www.ncgr.org/research/sequence/	包括所有核酸蛋白序列
Ensembl	http://www.ensembl.org/	注解的人类基因组序列数据
STACK	http://www.sanbi.ac.za/Dbases.html	非冗余基因簇
TIGR Gene Indices	http://www.tigr.org/tdb/index.html	非冗余基因簇
UniGene	http://www.ncbi.nlm.nih.gov/UniGene/	非冗余基因簇
比较基因组学		
Clusters of Orthologous Groups (COG)	http://www.ncbi.nlm.nih.gov/COG/	根据 44 种已完成基因组蛋白的进化树分类
Comparative Genometrics	http://www.unil.ch/igbm/genomics/genometrics.html	全基因组的生物统计学比较
euGenes	http://iubio.bio.indiana.edu:89/	真核生物数据库中基因和基因组信息的概括
Genome Information Broker	http://gib.genes.nig.ac.jp/	已完成的微生物基因组比较分析
Gramene	http://www.gramene.org/	禾本植物的比较基因组分析
Homophila	http://homophila.sdsc.edu/	人类疾病基因和果蝇基因的关系

数据库名称和分类	互联网网址	内容介绍
XREFdb	http://www.ncbi.nlm.nih.gov/XREFdb/	模式生物遗传学和哺乳动物表型的相互参考
基因表达		
ASDB	http://cbcg.nersc.gov/asdb	基因不同剪切方式的蛋白产物和表达谱
Axeldb	http://www.dkfz-heidelberg.de/abt0135/axeldb.htm	非洲蟾蜍的基因表达
BodyMap	http://bodymap.ims.u/	人和小鼠基因表达数据
EpoDB	http://www.cbil.upenn.edu/epodb/	脊椎动物红细胞的基因表达
EPConDB	http://www.cbil.upenn.edu/EPConDB	内分泌胰腺协议数据库
FlyView	http://pbio07.uni/	果蝇发育和遗传学
Gene Expression Database (GXD)	http://www.informatics.jax.org/searches/gxdindex_form.shtml	小鼠基因表达和基因组
Gene Expression Omnibus (GEO)	http://www.ncbi.nlm.nih.gov/geo	基因表达和杂交微阵列数据库
HugeIndex	http://www.hugeindex.org/	人类基因在正常组织中的表达水平
Interferon Stimulated Gene Database	http://www.lerner.ccf.org/labs/williams/xchip-html.cgi	干扰素诱导或处理后的基因表达
Kidney Development Database	http://www.ana.ed.ac.uk/anatomy/database/kidbase/kidhome.html	肾发育和基因表达
MAGEST	http://star.scl.kyoto-u.ac.jp/magest/	海鞘类(Halocynthia roretzi)基因表达
MethDB	http://www.methdb.de/	DNA 甲基化数据和特征谱
Mouse Atlas and Gene Expression Database	http://genex.hgu.mrc.ac.uk/	不同部位基因表达数据图谱
READ	http://read.gsc.riken.go.jp/READ/	RIKEN 表达阵列数据库
PEDB	http://chroma.mbt.washington.edu/PEDB/	正常和异常前列腺基因表达
RECODE	http://recode.genetics.utah.edu/	表达中具有程序性

数据库名称和分类	互联网网址	内容介绍
Stanford Microarray Database	http://genome-www.stanford.edu/microarray	来自 DNA 芯片实验的原始和标准化的数据
TRIPLES	http://ygac.med.yale.edu/triples/triples.htm	酵母转座子插入表型定位和表达
Tooth Development Database	http://bite-it.helsinki.fi/	口腔组织的基因表达
yMGV	http://www.transcriptome.ens.fr/ymgv/	酵母微阵列数据和开发工具
基因识别和结构		
AllGenes	http://www.allgenes.org/	人和小鼠基因索引，整和基因转录和蛋白注解
Ares Lab Intron Site	http://www.cse.ucsc.edu/research/compbio/yeast_introns.html	酵母染色体内含子剪切位点数据
AsMamDB	http://166.111.30.65/ASMAMDB.html	不同剪切方式的哺乳动物基因
COMPEL	http://compel.bionet.nsc.ru/	复合调节元件
CUTG	http://www.kazusa.or.jp/codon/	密码子语法表
DBTBS	http://elmo.ims.u-tokyo.ac.jp/dbtbs/	枯草杆菌结合因子和启动子
DBTSS	http://elmo.ims.u-tokyo.ac.jp/dbtss/	转录起始位点
EID	http://mcb.harvard.edu/gilbert/EID/	内含子外显子数据库
EPD	http://www.epd.isb-sib.ch/	真核生物 POL II 启动子和实验检测的转录起始位点
ExInt	http://intron.bic.nus.edu.sg/exint/exint.html	真核基因外显子内含子结构
HUNT	http://www.hri.co.jp/HUNT	已注解的全长 cDNA 序列
FUGOID	http://wnt.cc.utexas.edu/~ifmr530/introndata/main.htm	细胞器内含子的功能结构信息
Gene Resource Locator	http://grl.gi.k.u-tokyo.ac.jp/	用已完成的人类序列排列 ESTs
HS3D	http://www.sci.unisannio.it/docenti/rampone/	人类外显子、内含子和剪切区域
HvrBase	http://www.hvrbase.org/	灵长类 mtDNA 控制区序列
IDB/IEDB	http://nutmeg.bio.indiana.edu/intron/index.html	内含子序列与进化

数据库名称和分类	互联网网址	内容介绍
PALSdb	http://palsdb.ym.edu.tw/	公认的不同剪切位点
PLACE	http://www.dna.affrc.go.jp/htdocs/PLACE	植物顺式激活调节元件
PlantCARE	http://sphinx.rug.ac.be:8080/PlantCARE/index.htm	植物顺式激活调节元件
PromEC	http://bioinfo.md.huji.ac.il/marg/promec	大肠杆菌 mRNA 启动子与实验确定的转录起始位点
RRNDB	http://rrndb.cme.msu.edu/	原核核糖体 RNA 操纵子的变异
STRBase	http://www.cstl.nist.gov/div831/strbase/	DNA 短串联重复序列
TransCOMPEL	http://compel.bionet.nsc.ru/FunSite/CompelPatternSearch.html	真核生物基因转录调节元件
SpliceDB	http://genomic.sanger.ac.uk/spldb/SpliceDB.html	经典和非经典的哺乳动物剪切位点
TRRD	http://wwwmgs.bionet.nsc.ru/mgs/dbases/trrd4	真核基因转录调节区
TransTerm	http://uther.otago.ac.nz/Transterm.html	密码子语法起始和终止信号
VIDA	http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html	病毒基因组开放读框
WormBase	http://www.wormbase.org/	美丽线虫生物学指南
YIDB	http://www.EMBL-Heidelberg.DE/ExternalInfo/seraphin/yidb.html	酵母核和线粒体内含子序列
rSNP Guide	http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/	调节基因区域的单核苷酸多态性
遗传学和物理图谱		
DRESH	http://www.tigem.it/LOCAL/drosophila/dros.html	与果蝇突变基因同源的人 cDNA 克隆
G3-RH	http://www-shgc.stanford.edu/RH/	斯坦福大学 G3 和 TNG 放射杂交图谱
GB4-RH	http://www.sanger.ac.uk/Software/Rhserver/Rhservers.html	Genebridge4 (GB4) 人放射杂交图谱
GDB	http://www.gdb.org/	人类基因和基因组图谱
GenAtlas	http://www.citi2.fr/GENATLAS/	人类基因标记和表型
GenMapDB	http://genomics.med.upenn.edu/genmapdb	已作图的人类 BAC

数据库名称和分类	互联网网址	内容介绍
GeneMap '99	http://www.ncbi.nlm.nih.gov/genemap/	国际放射杂交协议人类基因图谱
HuGeMap	http://www.infobiogen.fr/services/Hugemap	人基因组遗传和物理图谱数据
IXDB	http://ixdb.mpimg-berlin-dahlem.mpg.de/	人类染色体 X 的物理图谱
RHdb	http://www.ebi.ac.uk/RHdb	放射杂交图谱数据
基因组数据库		
ACeDB	http://www.sanger.ac.uk/Software/Acedb/	美丽线虫、啤酒酵母和人的序列和基因组信息
AMmtDB	http://bio-www.ba.cnr.it:8000/BioWWW/#AMMTDB	后生动物线粒体 DNA 序列
Arabidopsis Information Resource (TAIR)	http://www.arabidopsis.org/	拟南芥基因组
ArkDB	http://www.thearkdb.org/genome_mapping.html	农场动物基因组数据库
Celera Discovery System	http://www.celera.com/genomics/academic/	整和的、以网络为基础的发现平台
Comprehensive Microbial Resource	http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl	已完成的微生物基因组数据
CropNet	http://ukcrop.net/	农业植物基因组作图
CyanoBase	http://www.kazusa.or.jp/cyano/	Synechocystis sp. 基因组
Dictyostelium Genome Sequencing Project	http://dictygenome.bcm.tmc.edu/	Dictyostelium 基因组资源
EcoGene	http://bmb.med.miami.edu/EcoGene/EcoWeb/	大肠杆菌 K-12 序列
EMGlib	http://pbil.univ-lyon1.fr/emglib/emglib.html	已完成测序的细菌、原生质和酵母基因组
FANTOM2	http://fantom.gsc.riken.go.jp/fantom2/doc/	RIKEN 小鼠基因百科全书计划 小鼠 cDNA 克隆的功能注解)
FlyBase	http://www.fruitfly.org/	果蝇序列和基因组信息

数据库名称和分类	互联网网址	内容介绍
Full-Malaria	http://133.11.149.55/	疟原虫红细胞期的全长cDNA 文库
GOBASE	http://megasun.bch.umontreal.ca/gobase/gobase.html	细胞器基因组数据库
GOLD	http://igweb.integratedgenomics.com/GOLD/	完成和正在进行的基因组计划信息
HERV	http://herv.img.cas.cz/	人类内源的反转录病毒
HIV Sequence Database	http://hiv-web.lanl.gov/	HIV RNA 序列
HOWDY	http://gdb.tokyo.jst.go.jp/HOWDY	整和人类基因组信息,部分来源于原始序列
Human BAC Ends Database	http://www.tigr.org/tdb/humgen/bac_end_search/bac_end_intro.html	非冗余性人类 BAC 克隆末端序列
ICB	http://www.mbio.co.jp/icb	应用蛋白编码确认和分类细菌
INE	http://rgp.dna.affrc.go.jp/giot/INE.html	水稻遗传,物理图谱和序列信息
MagnaportheDB	http://www.cals.ncsu.edu/fungal_genomics/mgdatabase/int.htm	水稻和真菌 Magnaporthe grisea 整和物理和遗传图谱
MatDB	http://mips.gsf.de/proj/thal/db/	拟南芥基因组原始数据
Medicago Genome Initiative (MGI)	https://xgi.ncgr.org/mgi	模式豆类生物 Medicago 的 ESTs、基因表达和蛋白组数据
MITOMAP	http://www.gen.emory.edu/mitomap.html	人线粒体基因组
MITOP	http://websvr.mips.biochem.mpg.de/proj/medgen/mitop	线粒体蛋白,基因和疾病
Mendel Database	http://jiio6.jic.bbsrc.ac.uk/	与植物基因家族数据相关的 EST 和 STS数据库
MitBASE	http://www3.ebi.ac.uk/Research/Mitbase/mitbase.pl	线粒体基因组,物种间的变异和突变
MitoDat	http://www-lecb.ncifcrf.gov/mitoDat/	线粒体蛋白,主要是人类蛋白
MitoNuc/MitoAln	http://bio-www.ba.cnr.it:8000/srs6/	编码线粒体蛋白的核基因
Mouse Genome Database (MGD)	http://www.informatics.jax.org/	小鼠遗传学和基因组学

数据库名称和分类	互联网网址	内容介绍
Munich Information Center for Protein Sequences (MIPS)	http://www.mips.biochem.mpg.de/	蛋白和基因组序列
NRSub	http://pbil.univ-lyon1.fr/nrsub/nrsub.html	B. subtilis 基因组
Oryzabase	http://www.shigen.nig.ac.jp/rice/oryzabase/	水稻遗传学和基因组学
PlasmoDB	http://plasmodb.org/	疟原虫基因组
Phytophthora Genome Consortium Database	https://xgi.ncgr.org/pgc	Phytophthora infestans 和 Phytophthora sojae 的ESTs
Proteome BioKnowledge Library	http://www.proteome.com/	模式生物致病原和哺乳动物蛋白组
Rat Genome Database	http://rgd.mcw.edu/	大鼠遗传学和基因组学数据
RiceGAAS	http://RiceGaas.dna.affrc.go.jp/	水稻基因组序列和预测的基因组结构
RsGDB	http://www-mmg.med.uth.tmc.edu/sphaeroides	Rhodobacter sphaeroides 基因组
Saccharomyces Genome Database (SGD)	http://genome-www.stanford.edu/Saccharomyces	啤酒酵母基因组
SubtiList	http://genolist.pasteur.fr/SubtiList/	B. subtilis 168 基因组
TIGR Microbial Database	http://www.tigr.org/tdb/mdb/mdbcomplete.html	微生物基因组和染色体
The Arabidopsis Information Resource(TAIR)	http://www.arabidopsis.org/	拟南芥基因组
Wanda	http://www.evolutionsbiologie.uni-konstanz.de/Wanda/	复制的鱼类基因
WILMA	http://www.came.sbg.ac.at/wilma/	C. elegans 注解
ZFIN	http://www.zfin.org/	石斑鱼基因组
ZmDB	http://zmdb.iastate.edu/	玉米基因组
分子间相互作用		
Biomolecular Interaction Network Database(BIND)	http://binddb.org/	分子相互作用,复合物和代谢途径

数据库名称和分类	互联网网址	内容介绍
DIP	http://dip.doe-mbi.ucla.edu/	蛋白-蛋白相互作用目录
DPInteract	http://arep.med.harvard.edu/dpinteract/	大肠杆菌 DNA 结合蛋白结合位点
Database of Ribosomal Crosslinks(DRC)	http://www.mpimg-berlin-dahlem.mpg.de/~ag_ribo/ag_brimacombe/drc/	核糖体交联数据
MHC-Peptide Interaction Database	http://surya.bic.nus.edu.sg/mpid	MHC I类和 II类分子—多肽复合物
代谢途径和细胞调节		
ENZYME	http://www.expasy.ch/enzyme/	酶的分类命名
EcoCyc	http://ecocyc.pangeasystems.com/ecocyc/	大肠杆菌 K-12基因组,基因产物和代谢途径
EpoDB	http://www.cbil.upenn.edu/EpoDB/	人类红细胞生成的基因表达
FlyNets	http://gifts.univ-mrs.fr/FlyNets/FlyNets_home_page.html	果蝇分子间相互作用
GeneNet	http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/	结构的正式描述和基因网络的功能组织
Klotho	http://www.ibc.wustl.edu/klotho/	生物复合物的收集和分类
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.ad.jp/kegg	代谢和调节途径
LIGAND	http://www.genome.ad.jp/dbget/ligand.html	酶配体底物和反应
MetaCyc	http://ecocyc.org/	多种生物的代谢途径和酶
PathDB	http://www.ncgr.org/pathdb	生物化学途径,化合物和代谢
RegulonDB	http://www.cifn.unam.mx/Computational_Biology/regulondb/	大肠杆菌转录调节和操纵子组织
UM-BBD	http://www.labmed.umn.edu/umbbd/	微生物生物催化反应和生物降解途径
WIT2	http://wit.mcs.anl.gov/WIT2/	代谢模型的功能治疗和发展整和系统

数据库名称和分类	互联网网址	内容介绍
突变数据库		
16S and 23S Ribosomal RNA Mutation Databases	http://ribosome.fandm.edu/	16S、23S 核糖体 RNA突变数据库
ALFRED	http://alfred.med.yale.edu/alfred/index.asp	等位基因频率和DNA多态性
Androgen Receptor Gene Mutations Database	http://www.mcgill.ca/androgendb/	雄激素受体基因的突变
Asthma Gene Database	http://cooke.gsf.de/asthmagen/main.cfm	哮喘和变态反应的连锁和突变遗传学
Asthma and Allergy Database	http://cooke.gsf.de/asthmagen/main.cfm	哮喘和变态反应的连锁和突变遗传学
Atlas of Genetics and Cytogenetics in Oncology and Haematology	http://www.infobiogen.fr/services/chromcancer/	肿瘤染色体异常
BTKbase	http://www.uta.fi/laitokset/imt/bioinfo/BTKbase/	X 染色体连锁的 γ 球蛋白血症突变记录
CASRDB	http://data.mch.mcgill.ca/casrdb/	导致家族性高钙低钙血症严重原发性新生儿高甲状旁腺素症和显性低钙血症的钙敏受体突变
Cytokine Gene Polymorphism Database	http://www. pam. bris. ac. uk/services/GAI/cytokine4.htm	细胞因子基因多态性,体内表达和疾病相关研究
Database of Germline p53 Mutations	http://www.lf2.cuni.cz/win/projects/germline_mut_p53.htm	人肿瘤和细胞系 p53 基因突变
dbSNP	http://www.ncbi.nlm.nih.gov/SNP/	单核苷酸多态性
DT40	http://genetics.hpi.uni-hamburg.de/dt40.html	鸡DT40 B 细胞基因敲除突变
FLAGdb/FST	http://genoplante-info.infobiogen.fr/	拟南芥 T-DNA 转化子
GRAP Mutant Databases	http://tinyGRAP.uit.no/GRAP/	家族性 G蛋白偶联受体突变
jSNP	http://snp.ims.u-tokyo.ac.jp/	日本人群的单核苷酸多态性

数据库名称和分类	互联网网址	内容介绍
HGVbase	http://hgvbase.cgr.ki.se/	基因序列多态性
HIV-RT	http://hivdb.stanford.edu/hiv/	HIV 反转录酶和蛋白酶序列变异
Haemophila B Mutation Database	http://www.umds.ac.uk/molgen/haemBdatabase.htm	因子 IX基因的点突变、短重复和缺失数据
Human Gene Mutation Database (HGMD)	http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html	人类遗传疾病相关的已知基因病变
Human PAX2 Allelic Variant Database	http://www.hgu.mrc.ac.uk/Softdata/PAX2/	人 PAX2 基因的突变
Human PAX6 Allelic Variant Database	http://www.hgu.mrc.ac.uk/Softdata/PAX6/	人 PAX6 基因的突变
Human Type I and Type III Collagen Mutation Database	http://www.le.ac.uk/genetics/collagen/	人 I 型和 III 型胶原基因突变
HvrBase	http://db.eva.mpg.de/Hvrbase/	灵长类 mtRNA 控制区序列
KMDB	http://mutview.dmb.med.keio.ac.jp/mutview3/kmeyedb/index.html	人类眼疾病基因的突变
KinMutBase	http://www.uta.fi/imt/bioinfo/KinMutBase/	致病的蛋白激酶突变
MmtDB	http://www.ba.cnr.it/~areamt08/MmtDBWWW.htm	原生质线粒体 DNA 的突变和多态性
Mutation Spectra Database	http://info.med.yale.edu/mutbase/	病毒、细菌、酵母和哺乳动物基因的突变
NCL Mutations	http://www.ucl.ac.uk/ncl/	神经原蜡样脂褐素(NCL)基因的突变和多态性
Online Mendelian Inheritance in Man	http://www.ncbi.nlm.nih.gov/Omim/	人类遗传和基因组疾病目录
PAHdb	http://www.mcgill.ca/pahdb/	苯丙氨酸羟化酶位点的突变
PHEXdb	http://data.mch.mcgill.ca/phexdb	导致 X 染色体连锁高磷血症的 PHEX 基因突变

数据库名称和分类	互联网网址	内容介绍
PMD	http://pmd.ddbj.nig.ac.jp/	蛋白突变数据汇编
PTCH1 Mutation Database	http://www.cybergene.se/PTCH/ptchbase.html	PTCH1突变和单核苷酸多态性
RB1 Gene Mutation Database	http://www.d-lohmann.de/Rb/	人RB1基因突变
Ribosomal RNA Mutational Database	http://ribosome.fandm.edu/	16S和23S核糖体RNA突变数据库
SV40 Large T-Antigen Mutant Database	http://bigdaddy.bio.pitt.edu/SV40/	SV40大T抗原基因的突变
iARC p53 Database	http://www.iarc.fr/p53/	文献报道的人p53基因错意突变和短缺失
p53 Databases	http://metalab.unc.edu/dnam/mainpage.html	人p53和hprt基因突变,啮齿类转基因动物lacI和lacZ突变
病理学		
AngioDB	http://angiodb.snu.ac.kr/	血管生成和血管生成相关分子数据库
FIMM	http://sdmc.krdl.org.sg:8080/fimm/	功能分子免疫学数据
HCForum	http://hcforum.imag.fr/welcome_eng.html	人类细胞遗传学数据库
IDR	http://www.uta.fi/imt/bioinfo/idr/	免疫缺陷突变
Mouse Tumor Biology Database (MTB)	http://tumor.informatics.jax.org/	小鼠肿瘤命名、分类、发病率、病理和遗传因子
Oral Cancer Gene Database	http://www.tumor-gene.org/Oral/oral.html	口腔肿瘤相关基因的细胞分子和生物学数据
PEDB	http://chroma.mbt.washington.edu/PEDB/	来自前列腺组织和特异细胞类型cDNA文库的序列
Tumor Gene Family Databases (TGDBs)	http://www.tumor-gene.org/tgdf.html	各种肿瘤相关基因的细胞分子和生物学数据
蛋白质数据库		
AARSDB	http://rose.man.poznan.pl/aars/index.html	tRNA 氨基转肽酶序列

数据库名称和分类	互联网网址	内容介绍
ABCdb	http://ir2lcb.cnrs-mrs.fr/ABCdb/	ATP结合蛋白超家族转运蛋白
AraC/XylS database	http://www.arac-xyls.org/	细菌中 AraC/XylS 家族的阳性调节子
ASPD	http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd	人工的蛋白质和多肽
BRENDA	http://www.brenda.uni-koeln.de/	酶的功能数据
CSDBase	http://www.chemie.uni-marburg.de/~csdbase	包含冷休克结构域的蛋白
DatA	http://luggagefast.Stanford.EDU/group/arabprotein/	已注解的拟南芥编码序列
DExH/D Family Database	http://www.columbia.edu/~ej67/dbhome.htm	DEAD-box，DEAH-box 和 DexH-box 蛋白
ESTHER	http://www.ensam.inra.fr/cholinesterase/	酯酶和 α/β 水解酶以及相关酶
Endogenous GPCR List	http://www.biomedcomp.com/GPCR.html	G蛋白偶联受体在细胞系中的表达
FUNPEP	http://www.gpcr.org/FUNPEP/db	低复杂性和复合倾向蛋白序列
EXProt	http://www.cmbi.nl/exprot	实验证实功能的蛋白
GPCRDB	http://swift.embl-heidelberg.de/7tm/	G蛋白偶联受体
GenProtEC	http://genprotec.mbl.edu/	大肠杆菌 K-12 基因组基因产物和同源序列
HIV Molecular Immunology Database	http://hiv-web.lanl.gov/immunology/	HIV 表位
HUGE	http://www.kazusa.or.jp/huge/	人类大分子蛋白(>50KD)和 cDNA 序列
Histone Database	http://genome.nhgri.nih.gov/histones/	组蛋白和组蛋白折叠序列和结构
Homeobox Page	http://copan.bioz.unibas.ch/homeo.html	与 Homeobox 蛋白分类和进化相关的信息
Homeodomain Resource	http://genome.nhgri.nih.gov/homeodomain	Homeodomain 序列结构和相关遗传和基因组信息

数据库名称和分类	互联网网址	内容介绍
IMGT	http://imgt.cines.fr:8104/	人类和其他脊椎动物免疫球蛋白、T细胞受体和MHC序列
IMGT/HLA	http://www.ebi.ac.uk/imgt/hla/	人类主要组织相容性复合体
InBase	http://www.neb.com/neb/inteins.html	Intervening 蛋白序列(inteins)和基序
Kabat Database	http://immuno.bme.nwu.edu/	免疫学相关的蛋白序列
LGICdb	http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html	配体开放的离子通道亚基序列
MEROPS	http://www.merops.co.uk/	蛋白水解酶(蛋白酶和肽酶)
MHCPEP	http://wehih.wehi.edu.au/mhcpep/	MHC结合肽
Membrane Protein Database	http://biophys.bio.tuat.ac.jp/ohshima/database/	膜蛋白序列跨膜区和结构
MetaFam	http://metafam.ahc.umn.edu/	整和蛋白家族信息
MHCBN	http://www.imtech.res.in/raghava/mhcbn/	MHC结合和非结合的多肽
Nuclear Receptor Resource	http://nrr.georgetown.edu/nrr/nrr.html	核受体超家族
NUREBASE	http://www.ens-lyon.fr/LBMC/laudet/nurebase.html	核内激素受体
Olfactory Receptor Database	http://ycmi.med.yale.edu/senselab/ordb/	嗅觉受体样分子序列
ooTFD	http://www.ifti.org/	转录因子和基因表达
PKR	http://pkr.sdsc.edu/	蛋白激酶序列、酶学遗传学和分子结构性质
PPMdb	http://sphinx.rug.ac.be:8080/ppmdb/index.html	拟南芥胞浆膜蛋白序列和表达数据
PROMISE	http://bioinf.leeds.ac.uk/promise/	蛋白活性位点的活化中心和金属离子
Peptaibol	http://www.cryst.bbk.ac.uk/peptaibol/welcome.html	抗生素多肽序列
PhosphoBase	http://www.cbs.dtu.dk/databases/PhosphoBase/	蛋白磷酸化位点
PLANT-PIs	http://bighost.area.ba.cnr.it/PLANT-PIs/	植物蛋白酶抑制剂

数据库名称和分类	互联网网址	内容介绍
PlantsP	http://plantsp.sdsc.edu/	植物蛋白激酶和蛋白磷酸酶
Prolysis	http://delphi.phys.univ-tours.fr/Prolysis/	蛋白酶自然和合成的蛋白酶抑制剂
Protein Information Resource(PIR)	http://pir.georgetown.edu/	全面注解的非冗余蛋白序列数据库
Ribonuclease P Database	http://www.mbio.ncsu.edu/RNaseP/home.html	Rnase P 序列、排列和结构
SENTRA	http://wit.mcs.anl.gov/WIT2/Sentra/HTML/sentra.html	感觉信号传导蛋白
S/MARt db	http://transfac.gbf.de/SMARtDB/	Scaffold/matrix 连接区
SWISS-PROT /TrEMBL	http://www.expasy.ch/sprot	蛋白质序列
TIGRFAMs	http://www.tigr.org/TIGRFAMs	功能已确认的蛋白质家族资源
TRANSFAC	http://transfac.gbf.de/TRANSFAC/index.html	转录因子和结合位点
VIDA	http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html	同源的病毒蛋白家族
Wnt Database	http://www.stanford.edu/~rnusse/wntwindow.html	Wnt 蛋白和表型
trEST, trGEN and Hits	http://hits.isb/-sib.ch	预测的蛋白序列
蛋白质序列基序		
BLOCKS	http://blocks.fhcrc.org/	蛋白家族的保守序列区
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	保守蛋白结构域的排列模型
CluSTr	http://www.ebi.ac.uk/clustr/	SWISS-PROT + TrEMBL 蛋白自动分类到相关功能组
InterPro	http://www.ebi.ac.uk/interpro/	蛋白家族结构域和位点的整和文件资源
O-GLYCBASE	http://www.cbs.dtu.dk/databases/OGLYCBASE/	糖蛋白和 O 连接的糖基化位点
PIR-ALN	http://www-nbrf.georgetown.edu/pirwww/dbinfo/piraln.html	蛋白序列排列
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/	分级基因家族指纹

数据库名称和分类	互联网网址	内容介绍
PROSITE	http://www.expasy.ch/prosite/	具有生物学意义的蛋白质特征和基序
Pfam	http://www.sanger.ac.uk/Software/Pfam/	多序列排列和共有蛋白结构域的 hidden Markov 模型
ProClass	http://pir.georgeown.edu/gfserver/proclass.html	由 PIR 超家族和 PROSITE 特征定义的蛋白家族
ProDom	http://www.toulouse.inra.fr/prodom.html	蛋白质结构域家族
ProtoMap	http://www.protomap.cs.huji.ac.il/	SWISS-PROT 蛋白的自动分级分类
SBASE	http://www3.icgeb.trieste.it/~sbasesrv/	注解的蛋白结构域序列
SMART	http://smart.embl-heidelberg.de/	信号结构域序列
SUPFAM	http://pauling.mbu.iisc.ernet.in/~supfam	与结构相关的序列排列
SYSTERS	http://www.dkfz-heidelberg.de/tbi/services/cluster/systersform	利用各种其他信息资源的注解将蛋白序列分类为分离簇
eMOTIF	http://motif.stanford.edu/emotif	蛋白序列基序确认和查找
iPROCLASS	http://pir.georgetown.edu/iproclass/	已注解的蛋白分类数据库
蛋白组资源		
Aaindex	http://www.genome.ad.jp/dbget/	多肽的物理化学性质
GELBANK	http://gelbank.anl.gov/	已完成基因组的 2D 凝胶电泳特征
Human Proteome Survey Database	http://www.proteome.com/services	人小鼠和大鼠蛋白组的详细信息
Predictome	http://predictome.bu.edu/	蛋白之间预测的功能联系
Proteome Analysis Database	http://www.ebi.ac.uk/proteome/	全基因组蛋白质的功能分类工具 interpro和 clustr 的在线应用
REBASE	http://rebase.neb.com/rebase/rebase.html	限制性内切酶和相关的甲基化酶
SWISS-2DPAGE	http://www.expasy.ch/ch2d/	已注解的 2 维凝胶电泳数据库
Yeast Proteome Database (YPD)	http://www.proteome.com/databases/index.html	啤酒酵母的蛋白组

数据库名称和分类	互联网网址	内容介绍
YPL	http://fstgal12. tu-graz. ac. at: 7777/pls/al12/ypl. htm	绿色荧光蛋白标签和共聚焦显微镜确定的酵母蛋白定位
RNA 序列		
16S and 23S rRNA Mutation Database	http://ribosome. fandm. edu/	16S 和 23S 核糖体 RNA 突变数据库
5S Ribosomal RNA Database	http://biobases. ibch. poznan. pl/5SData/	5S rRNA 序列
ACTIVITY	http://wwwmgs. bionet. nsc. ru/mgs/systems/activity/	功能性 DNA/RNA 位点活性
ARED	http://rc. kfshrc. edu. sa/	包含 AU 丰富元件的mRNAs
Collection of mRNA-like Noncoding RNAs	http://biobases. ibch. poznan. pl/ncRNA/	不编码蛋白的RNA 转录子
European Large Subunit Ribosomal RNA Database	http://rrna. uia. ac. be/lsu/index. html	利用二级结构信息排列核糖体 RNA 大亚基序列
European Small Subunit Ribosomal RNA Database	http://rrna. uia. ac. be/ssu/index. html	利用二级结构信息排列核糖体 RNA 小亚基序列
Guide RNA Database	http://www. biochem. mpg. de/~goeringe/	Guide RNA 序列
HyPaLib	http://bibiserv. techfak. uni-bielefeld. de/HyPa/	RNA 不同分类的结构元件特征
Intronerator	http://www. cse. ucsc. edu/~kent/intronerator/	线虫 RNA 剪切和结构 美丽线虫和 briggsae 线虫基因组序列的排列
Non-Canonical Interactions in RNA	http://prion. bchs. uh. edu/bp_type/	已知 RNA 结构的非标准碱基 碱基相互作用
PLMItRNA	http://bigarea. area. ba. cnr. it:8000/PLMItRNA/	光合成真核生物的线粒体 tRNA 基因和分子
Pseudobase	http://wwwbio. leidenuniv. nl/~Batenburg/PKB. html	RNA 假结点(pseudoknots)信息
RISCC	http://ulises. umh. es/RISSC	16S 和 23S 核糖体 RNA 基因间隔区

数据库名称和分类	互联网网址	内容介绍
RNA Modification Database	http://medlib.med.utah.edu/RNAmods/	RNA 自然修饰核苷
Ribosomal Database Project (RDP)	http://rdp.cme.msu.edu/	rRNA 序列、排列和进化
SELEXdb	http://wwwmgs.bionet.nsc.ru/mgs/systems/selex/	选择的 DNA/RNA 功能位点序列
SRPDB	http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html	信号识别颗粒 RNA、蛋白和受体序列
Small RNA Database	http://mbcr.bcm.tmc.edu/smallRNA	从原核和真核生物直接测序小 RNA 序列
The tmRNA Website	http://www.indiana.edu/~tmrna	tmRNA 序列、折叠和排列
UTRdb/UTRsite	http://bigarea.area.ba.cnr.it:8000/EmbIT/UTRHome/	真核生物 mRNA3'和 5'非翻译区和相关功能特征
Viroids and viroid-like RNAs	http://nt.ars-grin.gov/subviral/	类病毒和类病毒样 RNAs
Yeast snoRNA Database	http://www.bio.umass.edu/biochem/rna-sequence/Yeast_snoRNA_Database/snoRNA_Data Base.html	酵母小核仁 RNAs
tRNA Sequences	http://www.uni-bayreuth.de/departments/biochemie/trna/	tRNA 和 tRNA 基因序列
tmRDB	http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html	tmRNA (10Sa RNA)序列
检索系统和数据库结构		
KEYnet	http://www.ba.cnr.it/keynet.html	数据检索用基因和蛋白分级列表
TESS	http://www.cbil.upenn.edu/tess	转录元件查找系统
Virgil	http://www.infobiogen.fr/services/virgil	数据库相互链接
结构		
ASTRAL	http://astral.stanford.edu/	已知结构的结构域序列,选择的序列-结构对应亚群
BioImage	http://www-embl.bioimage.org/	多维生物学图象可搜索数据库
BioMagResBank	http://www.bmrb.wisc.edu/	蛋白、多肽和核酸的 NMR 成像结构数据

数据库名称和分类	互联网网址	内容介绍
CATH	http://www.biochem.ucl.ac.uk/bsm/cath/	蛋白结构域结构的分级分类
CE	http://cl.sdsc.edu/ce.html	计算和观察 3D 蛋白结构排列的资源
CKAAPs DB	http://cl.sdsc.edu/ckaap	序列不相似但结构相似的蛋白
CSD	http://www.ccdc.cam.ac.uk/prods/csd/csd.html	有机复合物和金属有机复合物晶体结构信息
Database of Macromolecular Movements	http://bioinfo.mbb.yale.edu/MolMovDB/	描述蛋白质和大分子运动,包括电影
Decoys 'R'Us	http://dd.stanford.edu/	基于序列数据由计算机产生的蛋白质构象
DSDBASE	http://www.ncbs.res.in/~faculty/mini/dsdbase/dsdbase.html	蛋白质中天然的和人工的二硫键
GTOP	http://spock.genes.nig.ac.jp/~genome/gtop-j.html	从基因组序列预测的蛋白质结构
HIC-Up	http://alpha2.bmc.uu.se/hicup/	小分子异源复合物的结构
HSSP	http://www.sander.ebi.ac.uk/hssp/	结构家族和排列,结构保守区和结构域
IMB Jena Image Library of Biological Macromolecules	http://www.imb-jena.de/IMAGE.html	生物高分子三维结构分析和可视化
ISSD	http://www.protein.bio.msu.su/issd/	整和序列和结构信息
LPFC	http://www-smi.stanford.edu/projects/helix/LPFC/	蛋白家族核心结构文库
MMDB	http://www.ncbi.nlm.nih.gov/Structure/	所有实验测定的三维结构,与 NCBI Entrez 链接
ModBase	http://pipe.rockefeller.edu/modbase	已注解比较蛋白结构模型
NDB	http://ndbserver.rutgers.edu/NDB/ndb.html	包含核酸的结构
NTDB	http://ntdb.chem.cuhk.edu.hk/	核酸的热力学数据
PALI	http://pauling.mbu.iisc.ernet.in/~pali	同源蛋白结构的进化树和排列

数据库名称和分类	互联网网址	内容介绍
PDB	http://www.rcsb.org/pdb/	由X线晶体衍射和NMR确定的结构数据
PDB-REPRDB	http://www.rwcp.or.jp/papia/	基于PDB记录的代表性蛋白链
PDBsum	http://www.biochem.ucl.ac.uk/bsm/pdbsum	PDB结构的摘要和分析
PRESAGE	http://presage.berkeley.edu/	实验和预测注解的蛋白结构
ProTherm	http://www.rtc.riken.go.jp/jouhou/protherm/protherm.html	野生型和突变型蛋白的热力学数据
RESID	http://www-nbrf.georgetown.edu/pirwww/dbinfo/resid.html	蛋白结构修饰
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	家族和结构蛋白关系
SCOR	http://scor.lbl.gov/	RNA结构的关系
SLoop	http://www-cryst.bioc.cam.ac.uk/～sloop/	蛋白环结构的分类
SUPERFAMILY	http://stash.mrc-lmb.cam.ac.uk/SUPERFAMILY/	蛋白质结构超家族的分配
转基因		
Cre Transgenic Database	http://www.mshri.on.ca/nagy/cre.htm	Cre转基因小鼠系
Transgenic/Targeted Mutation Database	http://tbase.jax.org/	转基因动物和靶向突变的信息
其他生物医学内容		
BAliBASE	http://www-igbmc.u-strasbg.fr/BioInfo/BaliBASE2/index.html	多序列对齐比较的基准数据库
DBcat	http://www.infobiogen.fr/services/dbcat/	数据库目录
DrugDB	http://pharminfo.com/drugdb/db_mnu.html	有药物活性的复合物,种类和商品名
END	http://www.ibc.wustl.edu/biognosis/agora_interface/html/agora_entrance.html	酶的学术名称
Global Image Database	http://www.gwer.ch/qv/gid/gid.htm	已注解的生物学图象
GlycoSuiteDB	http://www.glycosuite.com/	N-和O-连接的糖基结构和生物学资源信息

数据库名称和分类	互联网网址	内容介绍
HOX-PRO	http://www.mssm.edu/molbio/hoxpro/new/hox-pro00.html	Homeobox 基因簇
Imprinted Genes and Parent of Origin Effects	http://www.otago.ac.nz/IGC	动物印记基因和母本源性效应
LocusLink/RefSeq	http://www.ncbi.nlm.nih.gov/LocusLink/	遗传位点的序列和描述信息
MPDB	http://www.biotech.ist.unige.it/interlab/mpdb.html	证实作为引物和探针有用的合成寡核苷酸信息
Molecular Probe Database	http://srs.ebi.ac.uk/	合成寡核苷酸探针和 PCR 引物
NCBI Taxonomy Browser	http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html	在遗传数据库中至少有一个核酸或蛋白序列的所有生物的名称
PubMed	http://www.ncbi.nlm.nih.gov/PubMed/	MEDLINE 和 Pre-MEDLINE 引证信息
PharmGKB	http://pharmgkb.org/	由于人个体差异对药物反应的差异数据库
RIDOM	http://www.ridom.de/	基于 rRNA 序列确认医学微生物
SWEET-DB	http://www.dkfz-heidelberg.de/spec2/	注解的碳水化合物结构和物质信息
Therapeutic Target Database	http://xin.cz3.nus.edu.sg./group/ttd/ttd.asp	治疗用蛋白质和核酸靶位点、代谢途径和药物信息
Tree of Life	http://phylogeny.arizona.edu/tree/phylogeny.html	进化树和生物学多样性信息
Vectordb	http://vectordb.atcg.com/	核酸载体的特征和分类
VirOligo	http://viroligo.okstate.edu/	用于 PCR 和杂交的病毒特异性寡核苷酸

附录二　生物信息学定义一览表

机　构	给出的定义
美国乔治亚理工大学	生物信息学是采用数学、统计学和计算机等方法分析生物学、生物化学和生物物理学数据的一种综合学科。
美国密苏里大学	生物信息学是获取、存储和处理生物学信息的一门科学与技术。
美国加利福尼亚大学洛杉矶分校	生物信息学是对生物学信息和生物学系统内在结构的研究，它运用数学和计算机科学的分析理论和实用工具将分散的生物学数据联系起来。
Whatis. com网站	生物信息学是以加快生物学研究为目的而开发计算机数据库和算法的一门科学。
美国国立卫生院	生物信息学是研究、开发和应用计算机工具和方法扩展生物学、医学、行为科学和卫生数据的利用，包括数据的获取、存储、组织、检索、分析或可视化，即运用信息科学的原理和技术使大量的、分散的和复杂的生命科学数据更加明了(understandable)和有用(useful)。

索　引

后　　记

这部书稿完成于去年的9月底，其后又反复修改了四个月，可谓是三易其稿在编书改稿的过程中，我们对这门学问的发展变化之快及其网络资讯的丰富浩翰，颇有感触。因此，努力尝试着写出一部较为全面、系统且能够反映出最新的进展和相关讯息的读本，以增强本书的实用价值。

本书的各种数据资料是截至到2002年1月底的最新资讯。其中的绝大部分，我们做了反复的核对。由于不同的数据库公布最新统计结果的日期不同，我们以能查寻到的最近结果为准。书末附录的《分子生物学数据库一览表》也是今年1月最新的修定结果，希望能对大家有所帮助。

另外，书中有不少英文专有名词及缩写，大部分附有汉译，有些未予这是考虑到一些缩写已为人们所熟知，以不译为妥。为了读者查找方便，目录中的不少项是按原英语名称编排的，但正文中会有相应的中文解释。

尽管我们作了一些努力，但有道是“书不尽言，言不尽意”，这本小书只能论及生物信息学中很少的一部分。且鲁鱼亥豕之误难免，诚希读者诸君批评指正

陈寅恪先生有言：“士之读书治学，盖将以脱心志于俗谛之桎梏,真理因得以发扬”。

愿与各位共勉。

王哲　　王林　　刘刚

2002-02-02

（王哲博士之 e mail：wangzhe70@sohu.com）

序

生物信息学（Bioinformatics）是应用数理和信息科学的理论和方法研究生命现象，组织和分析日益剧增的生物信息数据库的一门新兴学科。它主要利用计算机、网络技术和不断发展的各种软件，研究遗传物质的载体 DNA 及其编码的功能大分子蛋白质，对逐日增多的序列和结构进行收集、整理、储存、发布、提取和加工，并从中分析和发现新的序列，从而不断揭示人体生理和病理过程的分子基础，为人类疾病的预防、诊断和治疗提供根本依据。实际上，生物信息学不仅已经成为生物医学、遗传学、农学等学科发展的强大动力，而且也为药物设计提供了有效途径。

随着人类基因组计划的不断发展，生物信息学的研究范围已从结构基因组学扩展到功能基因组学，随之又出现了进化基因组学。生物信息学的根本任务之一是发现新的基因、蛋白及其功能。生物信息学的特点是投资少，见效快，效益大，适合我国的现实条件。本书编著者是在生物信息学第一线工作的青年科学工作者，他们通过钻研与实践，已经基本掌握了如何从因特网上不断收集数据，并能进行分析、归类与重组，发现新线索、新现象和新规律，不仅发现并克隆了与肿瘤分化相关的新基因，并登录 GenBank，对有的新基因的功能也做了初步研究，并以此为基础获得了国家自然科学基金的资助。可贵的是，他们还把自己应用生物信息学的经验，在《生命科学》上介绍。为了加速我国生物信息学的不断快速发展，培养一批在数理、信息科学、计算机科学和分子生物学方面均有造诣的跨学科人才的任务十分迫切，愿本书能在这一方面发挥积极作用，吸引更多有志之士参与生物信息学研究，用不断发展的生物信息学推动我国生命科学的发展，发现更多具有我国自主知识产权的生物大分子，为我国科技创新做出贡献。

黄高昇

二〇〇二年三月于第四军医大学

序

苍宇时空无垠，科学前沿无涯。

近年来，随着分子生物学、人类基因组计划的快速发展，相应地产生了一门新兴的学科——生物信息学。它的出现是生命科学、计算机网络技术快速发展的必然结果，同时又对包括分子生物学、免疫学、神经科学在内的许多学科的发展起到了良好的促进作用。我们还欣喜地看到这门学科对科研思维、科学工作方法的扩展和改进都有助益。

但是，这毕竟是一门崭新的学科，有许多生物学研究者、临床工作者和青年学生对此不够了解。而在国内，系统地、深入浅出地介绍这方面知识的书籍很少见到。《生物信息学概论》一书的出版提供了极好的参考资料和学习读本，能够起到普及和提高的作用，使生命科学工作者受到这方面的训练和培养，使年轻学子易于掌握其基础知识和研究方法。

我校三位青年学者：王哲、王林、刘刚，近几年十分关注这一学科的发展。他们在完成各自研究课题的同时，悉心钻研，掌握了丰富的相关资讯。本书就是他们厚积薄发、大胆尝试之作。这是一本具有较高学术水平的参考书，它的出版无疑会对生物信息学的普及，以及生物学各领域的深入研究起到积极的推动作用。

我热忱地祝贺本书的出版，并向广大生命科学工作者，特别是青年学者推荐此书。

胡 蕴 玉

辛巳岁末

于第四军医大学

前　　言

近十年，由于分子生物学在基因排列和蛋白质识别的研究上取得了可喜的进步，也由于对生物体功能和结构关系深入研究的必需，载录有数十亿数据信息的各类数据库需要有一个强有力的分析工具，用来描述数据与生物学意义之间的关联，用来收集、归纳、研究各类生物信息。这一工具就是生物信息学——一门传统生物学与计算生物学的交叉学科。

它的出现一方面是生命科学自身发展的需求；另一方面，信息科学、计算机及网络技术也为它的发展提供了理论支持和操作的平台。二者的结合使得对生物数据的演算、组织归纳和分析成为可能，并最终构架出具有生物学意义的本质内容。

如今，从事这一学科的研究开发、管理维护以及教学培训的专门人员已为数不少；应用这一工具为自己的科研服务的人就更多了。大致地，可以将他们分为 Doer 和 User 两类。前者是生物信息学的专业人员，包括各种研究机构（诸如： NCBI）的从业人员、大学里本专业的教研人员等等。他们中间有信息科学、分子生物学、结构生物学、计算机及网络技术、数学等方面的研究人员。而后者则是生物信息学的服务对象，包括生物学、医学、药学等学科的相关研究者。他们利用已建立好的各类数据库中的信息为自己的研究服务，同时也可能成为数据库的提交者和充实者。这本书就是为 User 提供基本知识的读物。

生物信息学的一个特点是发展速度很快。今天在网络上看到的东西已经与一年前有所不同了。形式上的不同仅是一方面，而更为重要的是内容上的变化。因此，写这方面的专著，常有跟不上变化的感觉；写成的东西也常常沦为“an Old Link”，而显得实用性不强。

在生物信息学的早期发展中，其变化固然多端；但在相对成

熟之后，其主要的形式和入口亦随之稳定下来。相关的方法学已详细地制定出来了，国际著名的一些数据库将会长期地发展下去。很多人发现，对生物信息学的基本内容有了相当的了解之后，追逐相关数据库的不断变化、进展，是一件令人着迷的事情。而这本书将就生物信息学的基础知识和最新进展做一系统的介绍。

这不是一本关于基因和蛋白质分析的实用手册，而是介绍基本概念、基本方法和生物学数据库最新资讯的专著。对于那些初入门的 User，这本书将是很有助益的。另外，本书的写作亦未过分简单化。其中的实用资料和解释，为研究者提供了有用的信息和帮助。

一年前，同学小聚。谈古论今之时，亦未敢遗忘正统学业。众人均对生物信息学有兴趣：言其发展神速，言其已使分子生物学进入了新境界，言其对研究方法、工作思维有深刻地影响，等等。深谈入巷，遂有著述之意。

其后的写作立即陷入了辛苦的套路之中，时常深感已入 harmless drudges 之境。然砥砺有成，今事随人愿。

但收获之余，有遗珠之恨；欣喜之际，有憾事不已。唯愿读者不吝赐教，以利我等不断地对此学问有新的领悟。

最后，感谢各位：

医学管理　医学硕士　王东光先生
病理学　医学博士　郭华章先生
放射学　医学学士　汤志华先生
病理学　医学博士　冯骥良先生

他们为本书的完成，提供了丰富的资讯服务和有益的信心支持。

作者　谨识

2001年10月2日

于第四军医大学